National Assessment
Governing Board

Content Alignment Studies of the 2009 National Assessment of Educational Progress for Grade 12 Reading and Mathematics with SAT and ACCUPLACER Assessments of these Subjects

Submitted: November 24, 2010

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

CONFIDENTIAL AND PROPRIETARY

Comprehensive Report: Alignment of 2009 NAEP Grade 12 Mathematics and ACCUPLACER Mathematics Core Tests

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-09-C-0001.

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

WestEd

Table of Contents

Executive Summary i

I. Introduction 1
- Purpose 1
- Governing Board's Approach to Preparedness 1
- Assessment-to-Assessment Alignment 2
- Alignment Study 3
- Report Overview and Organization 5

II. Methodology 6
- Study Design Overview 6
- Standards and Representation of the Mathematics Content Domain 7
- Comparison of Critical Features of the Assessments 8
- Item Pool Selection and Assessment Design 12
- Alignment Definition Used in the Study 14
- Alignment Criteria Used in the Study 15
- Depth-of-Knowledge Levels Used in the Study 17
- Adjudication Discussions Implemented in the Study 18
- Alignment Procedure Implemented in the Study 19
- Decision Rules 24
- Participants 25
- Preparation, Materials, and Logistics 30
- Pilot Study: Lessons Learned 35

III. Alignment Results 38
- Reliability and Interrater Agreement 38
- DOK Levels of the NAEP Framework 40
- DOK Levels of the ACCUPLACER Framework 42
- DOK Levels of the Test Items 42
- Alignment Results by Sub-Study 42

IV. Panelists' Evaluations of the Process 78

V. Summary and Conclusions 87
- Summary of Overlap of Content Alignment 87
- Overall Conclusions 105

VI. Discussion and Recommendations on Study Design 108

VII. References 113

Appendices Part 1

Appendix A. Alignment Study Design Document A-1
Appendix B. Interim Report: Comparative Analysis of the Test Blueprints and Specifications for 2009 NAEP Grade 12 Mathematics and ACCUPLACER Mathematics B-1
Appendix C. Test Specifications and Frameworks Showing Inter-Panel Consensus Depth-of-Knowledge Values C-1
Appendix D. WestEd NAEP Alignment Institute April 12–16, 2010 Mathematics Panels Agenda D-1
Appendix E. Panelist Training Materials.....E-1
Appendix F. Questionnaires and Evaluation Forms F-1
Appendix G. Facilitator Training Materials..... G-1
Appendix H. WestEd NAEP Alignment Institute Security Protocol..... H-1
Appendix I. Panelists' Responses to Evaluation Forms I-1

Appendices Part 2: Confidential and Proprietary

Appendix I. Panelists' Responses to Evaluation Forms (continued).....I-18
Appendix J. WAT Reports: NAEP–NAEP Mathematics Panels.....J-1
Appendix K. WAT Reports: ACCUPLACER–NAEP Mathematics Panels K-1
Appendix L. WAT Reports: ACCUPLACER–ACCUPLACER Mathematics Panels.....L-1
Appendix M. WAT Reports: NAEP–ACCUPLACER Mathematics Panels.....M-1
Appendix N. Assessments to ACCUPLACER Debrief (Mathematics) Responses N-1
Appendix O. Assessments to NAEP Debrief (Mathematics) Responses O-1

List of Tables

Table 1. Comparison of the Critical Features of the NAEP Grade 12 Mathematics Assessment and the ACCUPLACER Mathematics Assessment 8

Table 2. Interrater Agreement of Panels by Sub-Study 39

Table 3. DOK Findings for the NAEP Mathematics Framework 40

Table 4. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework 43

Table 5. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework 43

Table 6. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework 44

Table 7. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework 44

Table 8. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework ... 51

Table 9. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework 53

Table 10. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework 53

Table 11. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework 54

Table 12. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework 55

Table 13. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework 61

Table 14. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework 63

Table 15. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework 63

Table 16. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework 64

Table 17. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework 64

Table 18. Summary of Attainment of Acceptable Alignment Level on Three Content Focus Criteria as Rated by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework 68

Table 19. Range of Depth of Knowledge of ACCUPLACER Items Aligned to the ACCUPLACER Framework 69

Table 20. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework 70

Table 21. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework 70

Table 22. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework 71

Table 23. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework 71

Table 24. Summary of Attainment of Acceptable Alignment Level on Three Content Focus Criteria as Rated by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework 76

Table 25. Range of Depth of Knowledge of NAEP Items Aligned to the ACCUPLACER Framework 77

Table 26. Panelist Responses to Day 1 Training and Process Evaluation Questionnaire..... 78

Table 27. Panelist Responses to Day 2 Training and Process Evaluation Questionnaire..... 80

Table 28. Panelist Responses to Day 3 Process Evaluation Questionnaire 81

Table 29. Panelist Responses to Day 4 Process Evaluation Questionnaire 82

Table 30. Panelist Responses to End-of-Study Evaluation Questionnaire 83

Table 31. Panelist Responses Regarding Adequacy of Facilities 85

Table 32. Summary of the Overlap of Content Alignment between NAEP and ACCUPLACER Items and the NAEP and ACCUPLACER Frameworks at the Standard Level 87

Table 33. Summary of the Overlap of Content Alignment between NAEP and ACCUPLACER Items and the NAEP Framework at the Objective Level..... 89

Table 34. Summary of the Overlap of Content Alignment between NAEP and ACCUPLACER Items and the ACCUPLACER Framework at the Objective Level..... 97

Acknowledgments

This study was funded by the National Assessment Governing Board under Contract ED-NAG-09-C-0001 and was managed by the Assessment and Standards Development Services (ASDS) program within WestEd.

Study Facilitators:
Michael Brown
Linda McQuillen

Study Panelists:
[redacted]

Technical Advisor:
Norman Webb

WestEd Staff:
Stanley Rabinowitz
Peter Worth
Jennae Bulat
Greg Hill, Jr.
Jennifer Verrier

Information in this report regarding the specifications for the ACCUPLACER Mathematics Core Tests is derived from data provided by the College Board. Copyright © 2006–2008. The College Board. All rights reserved. No further use of Data is permitted. www.collegeboard.com. Formatting and numbering were added by WestEd for use in this study.

Important Notice

The research presented in this report was conducted under a contract with the National Assessment Governing Board. This research project is part of a larger program of multiple research projects that are being conducted for the Governing Board and that will be completed at different points in time.

The purpose of this program of research is to provide, collectively, validity evidence in connection with statements that might be made in reports of the National Assessment of Educational Progress (NAEP) about the academic preparedness of 12th grade students in reading and mathematics for postsecondary education and training.

The findings and conclusions presented in this research report, by themselves, do not support statements about 12th grade student preparedness in relation to NAEP reading and mathematics results. Readers should not use the findings and conclusions in this report to draw conclusions or make inferences about the academic preparedness of 12th grade students.

Comprehensive Report:
Alignment of 2009 NAEP Grade 12 Mathematics and ACCUPLACER Mathematics Core Tests

Executive Summary

The National Assessment Governing Board (Governing Board) contracted WestEd to independently evaluate and report on the extent to which the grade 12 National Assessment of Educational Progress (NAEP) is aligned in content and complexity to the SAT and the ACCUPLACER assessments in reading and mathematics. This series of alignment studies is an important component of the Governing Board's research initiative concerning the use of the grade 12 NAEP to report and explain findings regarding students' preparedness for higher education and workplace training or entry. The alignment study discussed in this report—one of four comprehensive reports to be submitted to the Governing Board—evaluated the alignment between the NAEP and ACCUPLACER assessments in mathematics.

While a typical alignment study explores the alignment between an assessment and a set of standards, this study investigated the degree of alignment between two assessments, assessments that were developed from different frameworks for different purposes. To accomplish its alignment objectives, the Governing Board proposed the use of a bi-directional, multifaceted study design developed by Dr. Norman Webb. This design, as implemented in this current study, comprised a qualitative comparison of the NAEP mathematics framework and the ACCUPLACER mathematics core test specifications, conducted in early 2010, and a series of alignment activities designed to investigate the degree of alignment between the pairs of assessments and frameworks/specifications.

These alignment activities were performed over the course of an alignment workshop conducted the week of April 12–16, 2010, and comprised a series of four sub-studies to determine the degree of alignment between 1) the grade 12 NAEP and the NAEP mathematics framework, 2) the ACCUPLACER assessment and the ACCUPLACER mathematics framework, 3) the grade 12 NAEP and the ACCUPLACER mathematics framework, and 4) the ACCUPLACER assessment and the NAEP mathematics framework. This bi-directional design allowed for a baseline of alignment to be determined between each assessment and its own framework/specifications, which was important in interpreting the degree of cross-framework/specifications alignment. A short-version representative sample of items was used for the within-framework analyses (i.e., NAEP items to NAEP framework and ACCUPLACER items to ACCUPLACER framework). The complete NAEP item pool and one complete form of three ACCUPLACER mathematics core tests were analyzed for the cross-framework analyses (NAEP items to ACCUPLACER framework and ACCUPLACER items to NAEP framework, respectively). Alignment criteria used and reported on in this study included categorical concurrence, depth-of-knowledge consistency (and range of depth of knowledge), range-of-knowledge correspondence, and balance of representation.

This report addresses the following specific questions:

- What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by ACCUPLACER?
- To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on ACCUPLACER?
- Are there systematic differences in content and complexity between NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

Summary of Findings

The four sub-studies show the following findings regarding the degree of alignment between each of the two assessments and its own framework as well as between each of the two assessments and the other assessment's framework. The standards in each framework are listed below.

NAEP Framework Standards

1. "Number properties and operations,"
2. "Measurement,"
3. "Geometry,"
4. "Data analysis, statistics, and probability," and
5. "Algebra."

ACCUPLACER Framework Standards

A. "Arithmetic,"
B. "Elementary algebra," and
C. "College level math."

NAEP Assessment to NAEP Framework Alignment

The NAEP short-version items (42 items) were found to assess all of the five NAEP standards. Of these five, "Algebra" received the greatest number of item alignments. The "Number properties and operations," "Measurement," and "Geometry" standards each received somewhat fewer item alignments, while the "Data analysis, statistics, and probability" standard received the fewest alignments.

ACCUPLACER Assessment to NAEP Framework Alignment

With regard to alignment to the NAEP framework, slightly over half of the 105 ACCUPLACER items were found to align to the NAEP "Algebra" standard, with somewhat fewer items aligning to "Number properties and operations." The NAEP "Measurement," "Geometry," and "Data analysis, statistics, and probability" standards each received few ACCUPLACER item alignments.

ACCUPLACER Assessment to ACCUPLACER Framework Alignment

The ACCUPLACER short-version items (45 items) were found to assess all of the three ACCUPLACER standards. While the percentages of aligned items were distributed relatively evenly across the three standards, "Elementary algebra" received slightly more item alignments than did "Arithmetic" or "College level math."

NAEP Assessment to ACCUPLACER Framework Alignment

NAEP items from the complete item pool (164 items) were also found to assess all of the three ACCUPLACER standards. Twenty of the 164 NAEP items were found to not align to the ACCUPLACER framework by the majority of panelists. However, considering only those items that were codable, NAEP items were found to align with a relatively even distribution to the three ACCUPLACER standards.

Categorical Concurrence

Categorical concurrence is met for a standard if at least six items are aligned to that standard. For alignment to the NAEP framework, the 42 NAEP short-version items were found to meet the typical WAT threshold value of at least six items for all standards except "Data analysis, statistics, and probability." Categorical concurrence was not met for this standard, although it approached this threshold. The 105 ACCUPLACER items met categorical concurrence for "Number properties and operations" and "Algebra," but not for the other standards.

For alignment to the ACCUPLACER framework, the 45 ACCUPLACER short-version items were found to meet categorical concurrence for all standards. The 164 NAEP items also met categorical concurrence for all ACCUPLACER standards.

In reviewing whether the categorical concurrence threshold is met, it is important to consider the impact of the number of items in the analyzed set (i.e., the more items that are analyzed, the more likely it is that the criterion will be met). In this study, the ACCUPLACER item pool was approximately two-thirds the size of the NAEP item pool.[1]

Depth-of-Knowledge Consistency and Range of Depth of Knowledge

Depth-of-knowledge consistency for a standard is met if at least 50% of the items aligned to an objective in that standard are at or above the DOK level assigned to that objective. For alignment to the NAEP framework, the NAEP items were found to meet depth-of-knowledge consistency for all standards. The ACCUPLACER items also met depth-of-knowledge consistency for all NAEP standards.

For alignment to the ACCUPLACER framework, DOK was analyzed as range of depth of knowledge. NAEP items aligned to the ACCUPLACER framework were coded at DOK

[1] The College Board provided a total of 165 unique ACCUPLACER mathematics items, comprising two parallel forms, with some item overlap. Due to timing considerations discussed in Section II of this report, only one form of each of the three ACCUPLACER mathematics core tests was coded, for a total of 105 items.

Levels 1–3, with most of the items at DOK Level 2. Almost all ACCUPLACER items aligned to the framework were found to be at DOK Level 1 or Level 2.

Range-of-Knowledge Correspondence

Range-of-knowledge correspondence is met for a standard if 50% or more of the objectives in that standard have items aligned to them. For alignment to the NAEP framework, the NAEP short-version items did not meet the criteria for range of knowledge for any standard. In fact, no NAEP standard had more than 39% of its objectives hit, with "Geometry" having the most restricted range of knowledge. This result likely reflects the large number of objectives (130), relative to the number of items used in this sub-study (42). For the ACCUPLACER items, range of knowledge was met for "Number properties and operations" and "Algebra" but was not met for "Measurement," "Geometry," or "Data analysis, statistics, and probability."

For alignment to the ACCUPLACER framework, the NAEP items had a range of knowledge for "Elementary algebra" and "College level math"; while one panel found the range of knowledge for "Arithmetic" to not be met, the other found that it was weakly met. For the ACCUPLACER items, range of knowledge was met for "Elementary algebra" and was weakly met for "Arithmetic." Range of knowledge was not met for "College level math."

Balance of Representation

Balance of representation indicates whether the item alignments are balanced among those objectives receiving item alignments. It is important to review balance of representation in conjunction with categorical concurrence and range-of-knowledge correspondence, since the number of aligned items and the percentage of objectives aligned can impact the balance of representation. NAEP items met the typical balance of representation threshold for all standards in the NAEP framework. The ACCUPLACER items met the balance of representation for "Measurement" and "Data analysis, statistics, and probability" but not for "Number properties and operations." One panel found that the ACCUPLACER items did not meet the balance of representation threshold for "Geometry" and "Algebra," while the other panel found that the criteria for balance was met for "Geometry" and was weakly met for "Algebra."

In relation to the ACCUPLACER framework, the NAEP items met the balance of representation criterion for "College level math"; and weakly met it for "Elementary algebra." The NAEP items did not meet balance of representation for "Arithmetic." The ACCUPLACER items met the criteria for balance of representation for all three ACCUPLACER standards.

Overall Conclusions

The following conclusions regarding the alignment of the 2009 NAEP Grade 12 Mathematics assessment and the ACCUPLACER Mathematics Core Tests can be drawn from the results of this alignment study.

What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by ACCUPLACER?

The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra (included in NAEP's "Number properties and operations" and "Algebra" standards and in ACCUPLACER's "Arithmetic," "Elementary algebra," and "College level math" standards), although their respective degrees of alignment and focus in these subdomains vary. Whereas the NAEP items focus primarily on number or number operations and algebra content at the grade 12 level, with an emphasis on problem solving and application of concepts at that grade level, the ACCUPLACER items span a wider developmental and grade-level range (from basic to more advanced).

This difference in focus is consistent with the purposes of the two assessments and their frameworks. The NAEP objectives are written to describe assessable content for grade 12 mathematics; thus, the 130 objectives tend to address the skills and concepts specific to that grade. The purpose of ACCUPLACER is to help determine appropriate placement for an individual student, and so the 87 ACCUPLACER objectives are spread more broadly across grade levels and are intended to be more general.

To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on ACCUPLACER?

Regarding alignment to the NAEP framework, within the "Number properties and operations" and "Algebra" standards, NAEP items had broader overall coverage of the NAEP objectives than did ACCUPLACER. The 42 NAEP items (the short version used for within-framework alignment) aligned to 72 NAEP objectives, whereas the 105 ACCUPLACER items (one complete form of each of the three ACCUPLACER Mathematics Core tests) aligned to only 56 NAEP objectives, with 44% of the ACCUPLACER item alignments aligning to only three NAEP objectives (all in "Number properties and operations" and "Algebra"). These differences in breadth and emphasis between the two assessments were evident across all NAEP standards. For example, in each assessment, items were aligned to four NAEP "Algebra" objectives for which the other assessment had no alignments, reflecting differences in emphasis within that standard.

Regarding alignment to the ACCUPLACER framework, ACCUPLACER items in the short version of 45 items covered all three standards—"Arithmetic," "Elementary algebra," and "College level math"—with a relatively even distribution, although "College level math" had the lowest percentage of item alignments. NAEP items in the full pool of 164 items also covered "Arithmetic," "Elementary algebra," and "College level math," with a fairly even distribution of approximately one-third of NAEP codable items aligned to each standard, although "Elementary algebra" received somewhat fewer item alignments. Despite these differences in emphasis, however, considering only codable items, the percentages of alignments to each ACCUPLACER standard were relatively evenly distributed in both assessments and similar in distribution across assessments. At the objective level, the distribution of item alignments to objectives was relatively even on both tests, although each assessment was aligned to some objectives to which the other was not.

In summarizing cross-framework alignment, there was somewhat less even distribution of items than observed in within-framework alignment. The majority of items on each test were found to align to objectives on the other test. However, the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from "Number properties and operations" in the NAEP framework, and four of the five goals in "Algebra." Conversely, the NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives.

Are there systematic differences in content and complexity between NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives (20 were rated uncodable by the majority of panelists in each panel). For other NAEP items that were aligned to an ACCUPLACER objective, there were often parts of those items not addressed by the objective. These items were coded as aligned, since they do assess an ACCUPLACER objective, but parts of the items also cover other skills not included in the ACCUPLACER framework.

Regarding differences in alignment of complexity, the items from both tests that aligned to the NAEP standards met the typical depth-of-knowledge (DOK) consistency threshold; that is, the items assessed the objectives at or above the DOK level of the objective. The items from both tests that aligned to the ACCUPLACER standards had somewhat different ranges of DOK. The ACCUPLACER short-version items were divided fairly evenly between Level 1 and Level 2. The NAEP items aligned to the ACCUPLACER framework had a wider range of DOK, with items at Level 1, 2, and 3, and a greater emphasis on Level 2 than was in the ACCUPLACER items.

I. Introduction

Purpose

Preparing students for postsecondary success—in college, in the workplace, and/or in the military—is a fundamental objective of the K–12 educational system; refining processes by which postsecondary preparedness is measured and reported is, therefore, of central importance to entities, such as the National Assessment Governing Board (Governing Board), that are tasked with evaluating the progress of education within the United States. For over two decades, the Governing Board has guided the development and use of the National Assessment of Educational Progress (NAEP) in monitoring student achievement in the nation across time and content areas, and the Governing Board now looks to enhance NAEP's role and relevance by establishing NAEP's capacity to collect and report data that may be used to draw valid conclusions about the preparedness of 12th grade students for postsecondary activities. To this end, in 2007, the Governing Board convened a Technical Panel on 12th Grade Preparedness Research (Technical Panel) to recommend research and validity studies that could be used to enable NAEP to report on preparedness for college and for job training programs in the civilian and military sectors.

The Technical Panel's recommended multi-method approach (National Assessment Governing Board, 2009c) includes conducting content alignment studies; exploring statistical relationships with assessments and outcomes data in postsecondary education and civilian and military job training programs; conducting criterion-based judgmental standard setting activities; and administering national surveys of postsecondary educational institutions. As part of this multi-method approach, the Governing Board contracted WestEd to independently evaluate and report "the extent to which the grade 12 NAEP is aligned in content and complexity to the SAT and to the ACCUPLACER for the two assessments in reading and mathematics" (National Assessment Governing Board, 2009a, p. 3). These alignment studies will provide the Governing Board with information on the use of the grade 12 NAEP to report and explain findings regarding students' preparedness for higher education and entry/placement in job training courses, information that will serve as the groundwork for the Governing Board's subsequent research (e.g., establishing statistical relationships between NAEP and assessments that serve as measures of postsecondary preparedness). This report, one of four in this series of studies conducted by WestEd, describes the alignment between the 2009 National Assessment of Educational Progress grade 12 mathematics (NAEP) and the ACCUPLACER mathematics core tests in the content areas of Arithmetic, Elementary Algebra, and College Level Math (ACCUPLACER). Findings from the studies of the alignment between NAEP and ACCUPLACER Reading Comprehension, SAT Critical Reading, and SAT Mathematics are presented in separate reports (WestEd, 2010a, 2010b, 2010c).

Governing Board's Approach to Preparedness

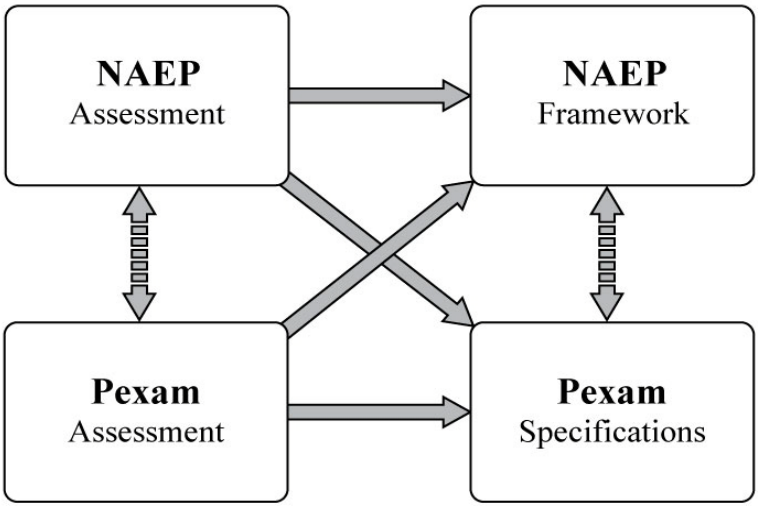
The Governing Board is focusing its conceptualization of 12th grade preparedness on academic qualifications and does not propose to address a range of behavioral and attitudinal aspects of student performance in postsecondary activities that are not measured by NAEP (e.g., time management skills, diligence). The Governing Board further limits its definition of postsecondary preparedness to refer to the academic skills required for placement into entry-level

college-level credit courses that count toward a four-year undergraduate degree, or for placement into military or civilian job training programs[2] (e.g., apprenticeship programs, vocational institute or certification programs, on-the-job training programs), with no prediction of success in such college-level courses or job training programs.

Assessment-to-Assessment Alignment

While a typical alignment study explores the alignment between an assessment and a set of standards, the Technical Panel called for studies that would investigate the degree to which NAEP is aligned in content and complexity to other assessments, assessments that were developed from different frameworks for different purposes. To accomplish this objective, the Governing Board contracted with Dr. Norman Webb to propose a bi-directional, multifaceted study design to look at alignment between an assessment and its own framework (e.g., NAEP with NAEP) and between an assessment and another assessment's framework or set of specifications (e.g., NAEP with ACCUPLACER), as illustrated in Figure 1 on the following page. (The full text of the resulting study design document is provided in Appendix A.) This study design comprises both a qualitative comparison of the NAEP mathematics framework with the ACCUPLACER mathematics specifications and a series of alignment activities to investigate the degree of alignment between the pairs of assessments and frameworks/specifications. The qualitative comparisons of each set of frameworks (comparative analyses) are used to inform expectations for alignment, raise potential alignment issues prior to item coding, and inform interpretations of the alignment results. This design is intended to ascertain the degree of alignment of two assessments by comparing how the items on the two assessments represent their respective content domains (National Assessment Governing Board, 2009b, p. 5).

Figure 1. Bi-Directional Alignment Methodology Overview[3]



This approach poses certain challenges, including the difficulty in standardizing the level at which analysis can occur across different content frameworks and the need to define and

[2] This conceptualization explicitly assumes that similar jobs in the military and civilian sectors require approximately similar academic skills and knowledge.
[3] In the design document, the term "Pexam" is the generic term used for the performance exams to which NAEP is compared in the series of alignment studies.

differentiate between constructs across the different frameworks. In addition, while many alignment studies investigate the overlap in content between an assessment and the framework upon which it was developed, or between an assessment and a set of standards to which the assessment was not originally developed, this approach was designed to align two assessments that were developed from different frameworks and for different purposes and uses.

Although both grade 12 NAEP and ACCUPLACER measure the mathematics skills of students at similar ages and stages of academic progress, they serve different purposes for different audiences. NAEP, commonly referred to as "the Nation's Report Card," is administered to representative samples of students across the country, and results are provided at the national level for grade 12. NAEP does not provide results for individual students. ACCUPLACER is primarily used by colleges and universities to help determine the appropriate placement of incoming students in college-level courses and "to determine if developmental classes would be beneficial before the students take college-level work" (College Board, 2009a). Therefore, ACCUPLACER provides results measuring the mathematics skills of individual students.

While a widely accepted standard of alignment for a typical alignment study may be a complete or nearly complete match between breadth and depth of content, the unique nature of this project and the differences that exist between the objectives and formats of the two assessments warrant modified expectations. As presented in Section III of this report, findings from this study are informed by the comparative analyses to most accurately contextualize the existing degree of alignment.

This report addresses the following specific questions:

- What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by ACCUPLACER?
- To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on ACCUPLACER?
- Are there systematic differences in content and complexity between the NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between the NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

Alignment Study

The NAEP–ACCUPLACER mathematics alignment study discussed in this report was conducted using the Governing Board's study design document developed for grade 12 NAEP alignment studies (National Assessment Governing Board, 2009b). The comparative analysis of the NAEP framework and ACCUPLACER specifications occurred in early 2010, while the alignment activities were performed over the course of an alignment workshop conducted the week of April 12–16, 2010, at the Westin Grand hotel in Washington, DC. The alignment study comprised a series of four sub-studies to determine the degree of alignment between 1) the grade 12 NAEP and the NAEP mathematics framework, 2) the ACCUPLACER assessment and the ACCUPLACER mathematics specifications, 3) the grade 12 NAEP and the ACCUPLACER mathematics specifications, and 4) the ACCUPLACER assessment and the NAEP mathematics

framework. This bi-directional design allowed for a baseline of alignment to be determined between each assessment and its own framework/specifications, which could be used in interpreting the degree of cross-framework/specifications alignment. Alignment criteria used and reported on in this study included categorical concurrence, depth-of-knowledge consistency, range of knowledge, and balance of representation.

The alignment workshop engaged two replicate panels of mathematics content experts, each comprising eight panelists, to independently and concurrently analyze assessment frameworks and assessment items. Each panel was led by an experienced group facilitator, with oversight provided by project management. Having two concurrent panels conduct the same analyses allowed for "a real-time check on the replicability (i.e., reliability) of the findings" (National Assessment Governing Board, 2009b, p. 10) and allowed for on-site adjudication and the real-time resolution of differences in interpretation. Descriptions of the expertise and training of the facilitators and panel members, as well as the means by which they were recruited, are provided in Section II of this report.

In order to capitalize on cost efficiencies, the NAEP–ACCUPLACER mathematics alignment study was conducted concurrently with the NAEP–ACCUPLACER reading alignment study also called for in this study's design document (National Assessment Governing Board, 2009b); as both studies occurred in the same meeting facility, WestEd staff and Governing Board representatives were able to oversee both studies simultaneously. This report describes only the results of the mathematics alignment study for these two assessments (see Section III of this report for alignment results).

The development of the NAEP mathematics framework document used in this study is described in Section II of this report; the resulting document is referred to in this report as the NAEP framework.[4] The development of the ACCUPLACER mathematics specifications document used in this study is also described in Section II of this report; the resulting document is referred to in this report as the ACCUPLACER framework.

[4] Concurrent with WestEd's alignment study, the Governing Board contracted with ACT for a separate study of the WorkKeys assessment using the same design document. To ensure consistency across the studies as appropriate, the Governing Board requested that WestEd and ACT share specific information and materials (e.g., NAEP reading framework organization, surveys, table formats, draft report of findings) developed during each other's studies, and facilitated conversations, including an in-person meeting, where issues of cross-project relevance (i.e., the NAEP framework, analysis methods, and reporting formats) were discussed. The sharing of information and materials was for the purpose of standardization of process and format and did not impact the content alignment judgments.

Report Overview and Organization

This report is organized as follows:

- Section II presents an overview of the methodology used to examine the alignment between the grade 12 NAEP and ACCUPLACER assessments in mathematics;
- Section III presents the results of this study;
- Section IV presents results of panelists' evaluations of the process;
- Section V presents a summary of results and conclusions;
- Section VI presents a discussion and recommendations regarding the study design;
- Section VII presents the references; and
- Appendices (Parts 1 and 2) conclude this report.

II. Methodology

This section begins with an overview of the components of the study design. This overview is followed by a detailed description of the methodology and study procedures; study participants; and preparation, materials, and logistics. The methodology, procedures, and logistics described in this section reflect lessons learned from the pilot alignment study of the NAEP and ACCUPLACER assessments in reading, which evaluated the appropriateness of the methodology, materials, and logistics as outlined in the study's design document (National Assessment Governing Board, 2009b) and as proposed by WestEd in this project's Planning Document. A summary of these lessons learned from the pilot study is provided at the end of the section.

Study Design Overview

This sub-section provides a high-level overview of the methodology implemented in this study. Each element of this study is described in greater detail later in this section.

This study implemented the study design document developed by Dr. Norman Webb for the Governing Board (National Assessment Governing Board, 2009b) to guide grade 12 NAEP alignment studies in evaluating the degree to which the grade 12 NAEP mathematics assessment aligns in content and complexity to the ACCUPLACER mathematics assessment. The study design called for a qualitative comparative analysis of the similarities and differences between the NAEP and ACCUPLACER frameworks. The result of this analysis is the NAEP–ACCUPLACER Interim Report, included as Appendix B.

Following the initial framework comparison, the study team implemented a content alignment workshop comprising a series of four sub-studies to determine the degree of alignment between 1) the grade 12 NAEP and the NAEP framework, 2) the ACCUPLACER assessment and the ACCUPLACER framework, 3) the grade 12 NAEP and the ACCUPLACER framework, and 4) the ACCUPLACER assessment and the NAEP framework. This bi-directional design allowed for a baseline of alignment to be determined between each assessment and its own framework (within-framework) as well as between each assessment and the other assessment's framework (cross-framework). The within-framework baseline alignment was important in interpreting the degree of cross-framework alignment.

The alignment methodology employed in this study called for each item to be assigned a DOK level and for each item to be coded to one primary and up to two secondary objectives, or to be rated "uncodable" if the item does not assess any objective. In addition, the methodology called for panelists to make note of items that contained source-of-challenge issues: items that students would either likely answer correctly without the intended knowledge or likely answer incorrectly despite having the intended knowledge.

The methodology also called for each objective within a standard to be coded to a DOK level. However, the pre-study review of the frameworks indicated that a modification to the study process was required for the ACCUPLACER mathematics framework. The ACCUPLACER mathematics framework is organized as a list of topics and does not provide sufficient information to determine the cognitive level at which the knowledge and skills in the objectives would be assessed. Without this information, it would have been impossible for panelists to

accurately code the objectives' DOK levels. Therefore, the ACCUPLACER framework was not coded for DOK. This step was replaced with a review of the ACCUPLACER objectives, during which panelists carefully reviewed the objectives to gain a level of familiarity with the framework approximating what they would have gain through DOK coding.

Over the course of the workshop, alignment coding occurred in the sequence indicated below:

1. NAEP framework reviewed and coded for DOK
2. NAEP items aligned to NAEP framework
3. ACCUPLACER framework reviewed
4. ACCUPLACER items aligned to ACCUPLACER framework
5. NAEP items aligned to ACCUPLACER framework
6. ACCUPLACER items aligned to NAEP framework

The Web Alignment Tool (WAT) was used to capture the alignment ratings of items and objectives and to analyze those ratings according to the Webb alignment criteria of categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. For alignment to the ACCUPLACER framework, depth-of-knowledge consistency was replaced by an analysis of the range of depth of knowledge of the aligned items.

Standards and Representation of the Mathematics Content Domain

The WAT system structure accommodates standards or frameworks that are structured hierarchically and that contain up to three levels. The three framework levels are labeled (in order of increasing specificity) as follows: standard, goal, and objective.

To assist in standardizing materials across the multiple alignment studies being conducted by the Governing Board, WestEd worked with the Governing Board, the project's technical advisor (Dr. Webb), and ACT to ensure that a NAEP mathematics framework organization appropriate for use in alignment studies was implemented. The form of the NAEP mathematics framework approved for this operational study was based on Exhibits 3–7 of the Governing Board's *Mathematics Framework for the 2009 National Assessment of Educational Progress* (National Assessment Governing Board, 2008, pp. 9–36), which present the mathematical content included in NAEP under five content areas: "Number properties and operations"; "Measurement"; "Geometry"; "Data analysis, statistics, and probability"; and "Algebra." Within each of these five content areas, the framework specifies subtopics and objectives at grades 4, 8, and 12. The content areas, subtopics, and grade 12 objectives were compiled into a single table, provided in Appendix C, organized into a three-tiered structure with 130 specific objectives at the most finely grained level. For use in the WAT, the five content areas were translated into standards (i.e., 1, 2, 3, 4, and 5). Within each standard, the subtopics were translated into goals (e.g., 1.1, 1.2, and 1.3). At the most specific level were the objectives (e.g., 1.1.d). The objectives in the original NAEP framework document are numbered and lettered consistently across grades, but not all objectives are appropriate for assessment at all grades. Therefore, not all letters appear in grade 12. For clarity and consistency with the original NAEP framework document, the numbering was kept consistent with the full framework. As such, there may appear to be some gaps in the numbering/lettering of the objectives (e.g., the first objective is 1.1.d).

The College Board provided a framework for each of the three ACCUPLACER mathematics core tests included in this study (i.e., Arithmetic, Elementary Algebra, and College Level Math). For each core test, the framework included a two-tiered description of test content, organized as broad topics and objectives. The three frameworks were combined for this study to create a single three-tiered set of standards, goals, and objectives, similar in structure to the NAEP framework previously described. One important distinction between the NAEP framework and the ACCUPLACER framework is that the ACCUPLACER framework indicates the content and skill topics but does not state the intended level of application of content and skill. This was identified in advance of the study and discussed with the Governing Board and the College Board. For the purposes of this study, the College Board provided an additional brief description to elucidate the intent of each category. However, as mentioned earlier in this section, the framework was determined to lack sufficient information on the intended level of application of skill for panelists to be able to code the objectives for depth of knowledge. As a result, it was determined that panelists could not code the ACCUPLACER framework for depth of knowledge. WestEd added alphanumeric coding to the framework corresponding to standard (e.g., A), goal (e.g., A.1), and objective (e.g., A.1.a) levels. The ACCUPLACER framework used in this study is included in Appendix C.

As discussed in greater depth in Section III of this report, alignment coding of items typically occurred at the objective level, although panelists were able to align an item to a goal or a standard if the item targeted no objectives.

Comparison of Critical Features of the Assessments

The full interim report comparing the content and structure of the assessment frameworks is included in Appendix B; Table 1 shows a comparison of the key features of the NAEP framework and the ACCUPLACER framework.

Table 1. Comparison of the Critical Features of the NAEP Grade 12 Mathematics Assessment and the ACCUPLACER Mathematics Assessment

	NAEP Grade 12 Math Assessment	**ACCUPLACER Math Assessment**
Percentage Distribution of Items by Content Area	Each NAEP mathematics item is developed to measure one of the objectives, which are organized into the four major content areas of mathematics: • Number Properties and Operations (10%) • Measurement and Geometry (30%) • Data Analysis, Statistics, and Probability (25%) • Algebra (35%)	Each ACCUPLACER mathematics core test is organized into major content areas, each of which contains 3–12 more specific subtopics (referred to in this study as “objectives”). For a given administration, the percentage of the test covered by each objective varies within a specified range of percentages of items. Some items meet multiple content requirements. Arithmetic Test: • Whole numbers and fractions • Decimals and percents • Applications Elementary Algebra Test: • Integers and rationals • Algebraic expressions • Equations, inequalities, and word problems College Level Math Test:

	NAEP Grade 12 Math Assessment	ACCUPLACER Math Assessment
		• Algebraic operations • Solutions of equations and inequalities • Coordinate geometry • Applications and other algebra topics • Functions • Trigonometry
Mathematical Complexity of Items	NAEP test takers spend the following percent of their testing time at each level of complexity: • Low (25%) • Moderate (50%) • High (25%)	No publicly available information, and none of the information furnished for this study describes the complexity of ACCUPLACER items, but general test levels include: • Arithmetic Test: assesses basic computation skills. • Elementary Algebra Test: assesses basic computation skills and the basic skills of algebraic manipulation that may be acquired in typical high school Algebra I and II courses. • College Level Math Test: assesses more advanced algebra skills typically required at the end of high school and beginning of college, as well as geometry and trigonometry.
Number of Items	The NAEP pool has 164 total mathematics items. No single student will complete all 164 items. Rather, each student completes two fixed item sets consisting of 13 or 14 items from the larger pool.	The ACCUPLACER computer-adaptive tests have the following numbers of items: • Arithmetic Test – 17 items • Elementary Algebra Test – 12 items • College Level Math Test – 20 items The ACCUPLACER "fixed form" version has 35 items.
Item Types	***Multiple choice*** 4 answer options: 1 correct, 3 incorrect ***Short constructed response*** 1- or 2-sentence response ***Extended constructed response*** 1- or 2-paragraph response	***All items are multiple choice*** 4 answer options: 1 correct, 3 incorrect
Time Per Item Type	The intended distribution of items for students is expressed as the percentage of time spent on each item type. • 50% multiple choice • 50% short and extended constructed response	Each test is untimed. Students can change answers to particular items before moving on to the next item, but cannot leave an item out or come back to it later to change answers.
Assessment Time	Each student will spend approximately 50 minutes (2 blocks at 25 minutes each) taking the NAEP assessment.	The test is untimed but designed to take less than one hour.
When Given	NAEP assesses and reports grade 12 mathematics results every four years.	ACCUPLACER administrations are determined by colleges and universities using the placement test.

	NAEP Grade 12 Math Assessment	ACCUPLACER Math Assessment
Testing Population	The 2009 Grade 12 NAEP was administered to: • 46,400 12th grade students in mathematics in 1500 public schools • Random samples of students designed to be representative of the nation • Samples of students in 11 states participating in a 2009 state-level pilot • ELL students unless they have had less than 3 school years of instruction in English • Students with disabilities unless their Individualized Education Plan (IEP) teams determine that they cannot participate, or whose cognitive functioning is so severely impaired that they cannot participate, or whose IEP requires an accommodation that NAEP does not allow	ACCUPLACER is administered to: • Students who are entering or planning to enter college at the freshman level
Accommodations	NAEP allows accommodations specified in an IEP that are routinely used in testing, such as: • Large-print material • Additional time • 1-on-1 or small-group testing • Having directions read • Preferential seating • Breaks during testing • Familiar person testing • Signing of directions • Signing of test items • Magnifying equipment • Template for response • Large marking pen or special writing tool for response • Pointing to answers or responding orally to transcribe Accommodations are offered in combination as needed; for example, students who receive one-on-one testing generally also use extended time. NAEP does not allow having items read aloud. For a complete list of NAEP math accommodations see: http://nces.ed.gov/nationsreportcard/about/inclusion.asp#accom_table	ACCUPLACER allows use of the following accommodations: • Recorded tests • Brailled versions of the tests • Large-print versions of the tests • Calculators • Interpreters, qualified readers, or transcribers • Screen display enlargement • Other effective methods of making orally delivered materials available to individuals with hearing impairments

	NAEP Grade 12 Math Assessment	ACCUPLACER Math Assessment
Calculator Use	The assessment contains blocks (sets of items) for which calculators are not allowed, and calculator blocks, which contain some items that would be difficult to solve without a calculator. Two-thirds of the blocks measure students' mathematical knowledge and skills without access to a calculator. One-third of the blocks allow use of a calculator. Students are allowed to bring any calculator they are accustomed to using in the classroom with some restrictions for test security purposes. Scientific calculators are supplied to students who do not bring a calculator to use on the assessment.	Calculators are permitted on some ACCUPLACER items. In the computer-based administration, a calculator icon indicates the availability of a pop-up four-function calculator for items on which calculator use is permitted. In the paper-and-pencil form, allowance of calculators is not recommended for the Arithmetic and Elementary Algebra tests. A four-function or scientific calculator may be used on the College Level Math test.

Item Scoring	The items are scored as:
	• Multiple choice:
	  • Incorrect 0
	  • Correct 1
	• Short constructed response:
	  • Incorrect 0
	  • Partial 1
	  • Correct 2
	• Extended constructed response:
	  • Incorrect 0
	  • Partial 1
	  • Essential 2
	  • Extensive 3
	All constructed-response items will be scored using rubrics unique to each item. General principles that apply to these rubrics follow:
	• Rubrics define minimal, partial, satisfactory, and extended responses.
	• Students will not receive credit for incorrect responses.
	• Student responses will be coded to distinguish between blank items and items answered incorrectly.
	• As part of the item review, the testing contractor will ensure a match between each item and the accompanying scoring guide.
		The items are scored as correct or incorrect.

	NAEP Grade 12 Math Assessment	ACCUPLACER Math Assessment
Test Scores	**Scaled scores**: Range of 0–300, average scores reported for groups **Achievement levels**: The numeric scale score range is divided into the following three achievement levels: • **Basic** — Partial mastery of prerequisite skills and knowledge necessary for proficient work • **Proficient** — Solid academic performance demonstrating competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter • **Advanced** — Superior performance Test scores and achievement levels are used to report on the performance of grade 12 students nationally. In 2009, 11 states participated in the first pilot for reporting state NAEP results at grade 12.	**Scaled scores**: Range of 20–120 ACCUPLACER provides results measuring the mathematics skills of individual students. Test scores are used to give college admissions and placement staff information about the academic readiness of students.

Item Pool Selection and Assessment Design

Selection of Item Pools for Alignment Workshop

The NAEP assessment design distributes the item pool across multiple test booklets using a matrix sampling design, so that a wider range of items can be assessed without burdening students. As a result, students taking the assessment will not all receive the same booklets or items. Each student completes two 25-minute timed item blocks with either 13 or 14 items in each block. The entire 2009 NAEP grade 12 mathematics item pool was included in this study. The item pool used consists of 164 items and includes multiple-choice items (1 point each) and constructed-response items (1 to 4 points each).

The ACCUPLACER mathematics assessment is a computer-adaptive test, consisting of a large pool of items from which a test-generation algorithm selects items for a student given that student's performance on prior items. All ACCUPLACER items are multiple choice and are dichotomously scored. Given the size of the total ACCUPLACER mathematics item pool, it would not be feasible to include all items in this study, even if the College Board had made the entire item pool available. More importantly, coding an entire adaptive item pool would not represent the assessment as administered. After extensive collaboration with WestEd and the Governing Board to determine the optimal item pool to use in this study, the College Board provided two paper-based forms (Forms F and G) that were developed for use by testing centers unable to administer the assessment via computer. These paper-based forms are an alternative format to the computer-adaptive administration, have been determined by the College Board to be representative of the ACCUPLACER item pool, have been used in other ACCUPLACER alignment studies, and were approved for use in this study by the Governing Board. Each paper-

based form consists of 35 items—20 items specific to that form (variable items) and 15 items common to both forms (common items). Considering all three core tests (i.e., Arithmetic, Elementary Algebra, and College Level Math), there were 105 items in Form F and 105 items in Form G, with an overlap of 45 common items. For purposes of efficiency and balance of workload across assessments and content areas, one of the two parallel forms made available for this study, Form F, was selected for use in this study, for a total of 105 items.

The study's design document (National Assessment Governing Board, 2009b) called for the entire item pool for each assessment to be aligned to both its own and the other assessment's framework; within-assessment alignment was conducted to provide a baseline level of alignment to inform interpretation of cross-assessment alignment ratings. However, based on WestEd pilot study experiences and lessons learned from the ACT mathematics alignment study for NAEP and WorkKeys, as well as the per-item time estimates provided in the design document, a modification was required. Given the large number of test items and content objectives, it was determined by WestEd and the Governing Board that there existed a substantial risk of not completing all alignment activities within the allotted time if the entire item pools were analyzed in each sub-study. The study was planned for five days, and it was determined to be unadvisable and a possible deterrent to recruiting to hold a workshop for longer than five days. In order to ensure that all alignment activities could be completed, WestEd and the Governing Board reached the solution of using a representative sample for alignment in the within-framework analyses. The reduction in data that would occur from using a sample set for the within-framework analysis was considered sufficient to meet the needs of the study (producing baseline alignment data and providing panelists exposure to each test's items in relation to its own framework) and preferable to not completing the study or having to reconvene panels at a later date. Therefore, with agreement by the study's technical advisor and author of the design document, WestEd and the Governing Board decided to limit the item pools as follows:

NAEP-to-NAEP Alignment

Following review of the entire NAEP item pool, WestEd recommended that a subset ("short version") consisting of 42 NAEP items be analyzed for alignment to the NAEP framework, with the goal of including the maximum number of items that could be analyzed during the planned coding time. The Governing Board concurred that using a short-version item pool would be sufficient if the items selected were representative of the total NAEP item pool. Following a review of the item pool and using the item-level characteristics provided for the NAEP items, WestEd selected a set of 42 items that would be representative of the range of items in the full item pool. This number was selected as large enough to be sufficiently representative of the full pool while small enough to allow for completion of the coding activities. The resulting short version sample item pool was a reasonable approximation of a representative sample, balancing the number of items with the following characteristics:

- mathematics content area (standard);
- complexity (high, moderate, or low);
- item type (multiple choice or constructed response);
- tool use (e.g., calculator, protractor, and/or ruler); and
- shared set leader, or common stimulus (e.g., items associated with the same figure or table).

In practice, efforts to balance these characteristics and include as many items as could be analyzed during the scheduled study produced a short version sample item pool that was within 1 percentage point of the total item pool distribution across the mathematics content areas of measurement, geometry, and algebra, but was overrepresentative of number, properties, and operations by approximately 9 percentage points (4–5 items), and underrepresentative of data analysis, statistics, and probability by approximately 12 percentage points (5 items). However, taking all the factors into account, this pool was considered by WestEd to represent a sufficient range of content, complexity, and the other item characteristics for use in the within-framework analysis.

NAEP-to-ACCUPLACER Alignment
The entire NAEP item pool was analyzed for alignment to the ACCUPLACER framework.

ACCUPLACER-to-ACCUPLACER Alignment
To reduce coding time given scheduling constraints, a subset ("short version") consisting of 45 items was analyzed for alignment to the ACCUPLACER framework. The 45 items comprised the 15 common (between Forms F and G) items from each of the three core tests. The common items were reviewed by the lead mathematics facilitator to ensure that they represented a range of content (standard, goal, and objective) and complexity. Additionally, WestEd compared the item difficulty statistics provided by the College Board (b-value) of the common items with those of the complete form and found them comparable.

ACCUPLACER-to-NAEP Alignment
The complete set of the 105 Form F items was analyzed for alignment to the NAEP framework.

For alignment purposes, within the WAT system, NAEP items were numbered sequentially in the order of their original block number, with the short version of 42 items listed first, followed by the remaining 122 items, also in block order, with items appearing in the order they were received from the National Center for Education Statistics (NCES). Within the WAT system, the ACCUPLACER Form F common items were numbered sequentially in the order in which they appear in the core test forms (15 each of Arithmetic, Elementary Algebra, and College Level Math), followed by the remaining 20 variable items per core test, numbered sequentially in the order in which they appear in the test form.

Alignment Definition Used in the Study

As described in this study's design document, alignment "generally attends to the agreement in content between state curriculum standards and state assessment. In general, two or more documents have content alignment if they support and serve student attainment of the same ends or learning outcomes. More specifically, alignment is the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (National Assessment Governing Board, 2009b, p. 2).

This study is different, however, in that—while a typical alignment study explores the alignment between an assessment and a set of standards—it attempts to investigate the degree to which two assessments align to each other, assessments that were developed from different frameworks for different purposes. As described earlier, to accomplish this objective, the Governing Board

proposed a bi-directional, multifaceted study design to look at within-framework alignment (e.g., NAEP with NAEP) and cross-framework alignment (e.g., NAEP with ACCUPLACER), and, in so doing, evaluate the degree of alignment of two assessments by comparing how the items on the two assessments represent their respective content domains.

Nevertheless, it is important to keep in mind that "alignment is an attribute of the relationship between two or more documents and less an attribute of any one of the documents. The alignment between a set of curriculum standards and an assessment could be improved by changing the standards, the assessment, or both" (National Assessment Governing Board, 2009b, p. 2). Particularly in a study of this nature, in which two documents developed in isolation from each other are compared, it is useful to take into consideration the unique characteristics and intended uses of each assessment when interpreting alignment results.

Alignment Criteria Used in the Study

The alignment methodology employed in this study used four criteria to determine the degree of alignment between the NAEP and ACCUPLACER assessments and the NAEP and ACCUPLACER frameworks, as defined by Dr. Webb.

Categorical Concurrence

"An important aspect of alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment, if both documents incorporate the same content. *The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents.* This criterion was judged by determining whether the assessment included items measuring content from each standard" (Webb, 2005, p. 110). For the purposes of this study, the typical WAT threshold value of six or more items had to target a given standard for the level of categorical concurrence between the standard and the assessment to be considered acceptable (indicated by a "Yes" in WAT reports). A "Weak" categorical concurrence rating was given by the WAT if five items were found to target a standard, while a "No" rating was given if four or fewer items were found to target a standard. Because the item counts vary greatly across the sub-studies, percentages of total hits and percentages of total hits adjusted for uncodable items also are provided in the report in order to facilitate comparisons across assessments.

Depth-of-Knowledge Consistency

"Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards*" (Webb, 2005, p. 111). For the purposes of this study, if 50% or more of items targeting a given standard were at or above the DOK level of the objective to which they aligned, that standard was given a "Yes" depth-of-knowledge consistency rating. If between 40% and 50% of items targeting a given standard were at or above the DOK level of the objectives to which they aligned, that standard was given a "Weak" depth-of-knowledge consistency alignment rating. A WAT rating of "No" depth-of-knowledge consistency indicated

that fewer than 40% of items targeting a standard were at or above the DOK level of the objectives to which they aligned.

As mentioned previously, the ACCUPLACER framework is organized as a list of topics and lacks sufficient information about the cognitive level of the knowledge skills to be coded for DOK; therefore, range of depth of knowledge analyses were conducted instead of depth-of-knowledge consistency analyses for alignment to the ACCUPLACER framework. This analysis examined the range of DOK levels assigned to the items aligned to each standard and may be a useful lens for examining alignment in the absence of DOK information on the framework.

Range-of-Knowledge Correspondence

"For standards and assessments to be aligned, the breadth of knowledge required on both should be comparable. The range of knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities. The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of objectives within the standard with one related assessment item/activity" (Webb, 2005, p. 112). For the purposes of this study, at least 50% of the objectives for a standard had to have at least one item aligned to them for the standard to be judged as having an acceptable range-of-knowledge correspondence. Particularly in studies such as this, in which item pools of substantially different sizes and frameworks of substantially different specificity are evaluated, it is important to note that this criterion is sensitive to the number of items being aligned and the level of detail of the frameworks to which they are being aligned, including the organization and number of standards, goals, and objectives.

Balance of Representation

"In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range of knowledge criterion only considers the number of objectives within a standard hit (a standard with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among these objectives. *The balance-of-representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another"* (Webb, 2005, p. 112). Typically, an index is used to judge the distribution of assessment items: "an index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a standard are equally distributed among the objectives for the given standard. Index values that approach 0 signify that a large proportion of the hits are on only one or two of all of the objectives hit" (Webb, 2005, p. 112). For the purposes of this study, an index value of 0.7 or higher was considered an acceptable balance of representation (represented by a "Yes" rating in the WAT), while an index value of 0.6 to 0.7 was considered a "Weak" alignment and an index value below 0.6 was considered to represent a lack of alignment (represented by a "No" rating in the WAT). These are the typical WAT threshold values. If an assessment's specifications call for a distribution that emphasizes particular objectives within a standard, that should be considered in reviewing the balance of representation index.

NAEP and ACCUPLACER will be compared through examining the attainment of the alignment criteria across the sub-studies.

Depth-of-Knowledge Levels Used in the Study

Four depth-of-knowledge levels were used to evaluate NAEP and ACCUPLACER assessments as well as the NAEP framework; they are described as follows:

> *Level 1 (Recall)* includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels, depending on what is to be described and explained.
>
> *Level 2 (Skill/Concept)* includes the engagement of some mental processing beyond an habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of objects or phenomena and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret," could be classified at different levels depending on the object of the action. For example, interpreting information from a simple graph, or reading information from the graph, also are at Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Level 2 activities are not limited only to number skills, but may involve visualization skills and probability skills. Other Level 2 activities include noticing or describing non-trivial patterns, explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.
>
> *Level 3 (Strategic Thinking)* requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3.
>
> Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and deciding which concepts to apply in order to solve a complex problem.

> *Level 4 (Extended Thinking)* requires complex reasoning, planning, developing, and thinking, most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include designing *and* conducting experiments and projects; developing and proving conjectures, making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs. (Webb, 2005, pp. 60–61)

Due to the focus in the Level 4 definition on higher-order thinking tasks carried out over an extended time period, panelists were trained that Level 4 could only apply to tasks (objectives or items) in which both higher-order thinking and extended time were factors, effectively excluding DOK Level 4 as an option for either NAEP or ACCUPLACER tasks.

Adjudication Discussions Implemented in the Study

In accordance with the replicate panel study design, adjudication discussions were held at scheduled points of the alignment process.

Adjudication of DOK of Objectives

As directed by the study's design document (National Assessment Governing Board, 2009b, p. 13), both mathematics panels were required to reach joint agreement on the DOK levels of each assessment framework's objectives[5]. As indicated earlier, the ACCUPLACER objectives were not coded for DOK; therefore, adjudication of DOK of objectives occurred only for the NAEP framework. Prior to alignment coding of the NAEP items, each panel independently coded the NAEP framework for DOK. Once coding was complete, the two panels individually adjudicated to achieve within-panel agreement on DOK levels; the facilitators then met separately to identify and adjudicate differences between the two groups to achieve cross-panel agreement on DOK levels. Upon reaching cross-panel agreement, the facilitators communicated these values to their panelists and entered NAEP framework objectives' DOK values into the WAT. In addition to providing important study data, the DOK adjudication process served a training and calibration purpose, ensuring that panelists were interpreting DOK consistently. Prior to alignment coding of ACCUPLACER items, each panel independently reviewed the ACCUPLACER objectives to gain familiarity with them. As the system used for data entry and analysis required a DOK value

[5] As stated in the design document regarding DOK coding of objectives, "Reaching true consensus among panel members is an important goal because the process affords the panel members the opportunity to discuss the fine points for each objective/element/skill" (National Assessment Governing Board, 2009b, p. 13).

to be entered for each objective, all ACCUPLACER objectives were assigned a default DOK level of 2.

Adjudication of DOK of Items and Alignment of Items to Frameworks

Both within-panel discussions and cross-panel adjudication sessions were held to discuss discrepancies in the coding of items to frameworks.

Within-Panel Discussion
After the panelists mapped items to an assessment framework, each facilitator reviewed her/his panelists' codes to ensure consistency of calibration and identify discrepancies in coding within the panel. Discrepancies that were identified for discussion included items that were assigned to three different DOK levels or to two non-contiguous DOK levels, and/or items that were not assigned to the same objective by more than half of the panelists. Discrepant items were then adjudicated within each panel, with the explicit instruction that panelists were not required to reach consensus, and panelists entered changes to their codes if their judgment of the coding had changed. This discussion of items with discrepant codes was to determine whether differences were based on a misinterpretation or systematic difference in application of the protocol, were related to specific issues with an item or standard, or were random differences among panelists.

Cross-Panel Adjudication
The facilitators then met separately with WestEd project staff and, usually, the Governing Board Contracting Officer's Representative (COR), to compare the results of the two groups for discrepancies as outlined in the design document. The facilitators and WestEd project staff reviewed the four alignment criteria—categorical concurrence (reviewing average numbers of items assigned to each objective), depth-of-knowledge consistency (reviewing average percentages of items at, below, and above the DOK level of the assigned objective)[6], range-of-knowledge correspondence (reviewing the percentages of objectives with at least one aligned item), and balance of representation (reviewing index values)—and discussed relevant items to determine whether the difference in coding was reasonable (i.e., not an error), and whether it was random or the result of a systematic difference in interpretation. Facilitators then reported back the outcomes of the cross-panel adjudication (i.e., areas of discrepancy, if any, and whether those discrepancies were systematic or random) to their respective panels, including raising specific items for discussion if necessary. Then, panelists were given the opportunity to change alignment codes based on the discussion.

Alignment Procedure Implemented in the Study

This alignment workshop occurred over five consecutive days. A full agenda by day is provided in Appendix D, although a summary of activities is included here to provide context for the discussion in Section III. As shown in the agenda, breakfasts and lunches were provided each day in order to accommodate an aggressive schedule, with the timing of morning and afternoon breaks determined by panel facilitators to coincide with natural stopping points in the work. Throughout the week, the two mathematics panels worked independently, with the facilitators meeting regularly to discuss progress and decision rules, and to identify items to be discussed

[6] For alignment to the ACCUPLACER framework, the range of depth of knowledge of item alignments was reviewed.

during within- and cross-panel adjudication; during most coding sessions and all adjudication sessions, a WestEd staff member was present to monitor and assist as needed.

To ensure that all groups received consistent information regarding the context of the overall study and the alignment methodology (e.g., use of replicate panels, purpose of adjudication discussions) and alignment criteria to be used in the study, both reading panels and both mathematics panels convened for an introductory session the morning of the first day, during which the project director provided an overview of the study's objectives, the study design, and definitions of the alignment criteria to be used in the alignment workshop; the COR provided an overview of the Governing Board, its mission, the NAEP assessment, and the preparedness research program; and a representative from the College Board provided an overview of the ACCUPLACER assessment. A copy of the PowerPoint presentation shared during this introductory session can be found in Appendix E. Following this introductory session, panels from the two content areas separated; for the remainder of the week, they reconvened as a whole group only for daily announcements prior to the start of each day's alignment activities, if necessary.

Following the introductory session, the two mathematics facilitators provided more detailed training in assigning DOK values to objectives to the combined mathematics panels. This initial training included group discussion of mathematics DOK levels and both group and individual practice coding sample objectives drawn from the *WAT Training Manual* (Webb, 2005). When the facilitators determined that the panelists were sufficiently calibrated in their understanding of DOK to begin assigning codes to the frameworks, the panelists separated into their individual panel rooms to register in the WAT. At the end of the first day of the alignment workshop, panelists were given the opportunity to indicate their levels of satisfaction with the training process via an online process evaluation questionnaire (provided in Appendix F).

As specified in the design document developed for this project, through the remainder of the week, each panelist independently performed the alignment tasks described below (see the study's design document, provided in Appendix A, for a detailed description of each, and see Appendix D for the schedule by which these tasks were conducted). Throughout the week, prior to beginning a new task or after an extended break, facilitators took a few moments to remind panelists of the criteria and tasks at hand.

Review NAEP Framework and Assign DOK Levels to Each Objective

Each panelist independently coded the NAEP framework for depth of knowledge. Once coding was complete, the two panels individually adjudicated to achieve within-panel agreement on DOK levels; the facilitators then met separately to identify and adjudicate differences between the two groups to achieve cross-panel agreement on DOK levels of the objectives. Upon reaching cross-panel agreement, the facilitators communicated the agreed-upon DOK values to their panelists and entered DOK values for the NAEP framework objectives into the WAT. In addition to providing important study data, the DOK adjudication process served a training and calibration purpose, in ensuring that panelists were interpreting DOK consistently.

Map NAEP Items to the NAEP Framework

Prior to mapping NAEP items to the NAEP framework, the combined mathematics panels convened to be trained in assigning DOK levels to items and mapping items to the NAEP framework. This training included a review of mathematics DOK levels and both group and individual coding of sample NAEP and ACCUPLACER assessment items[7]. Once the facilitators deemed the panelists to be sufficiently calibrated in coding items for both DOK levels and alignment to objectives, the panelists separated into their individual panel rooms. In each group, the facilitator led the panelists through the coding of a limited sample set of active NAEP items[8] from the item booklet, to ensure understanding of the task and calibration among panelists. As indicated earlier, a subset of 42 NAEP items was selected to be mapped to the NAEP framework; once calibration was reached, panelists began to independently map the remaining NAEP items from this 42-item subset to the NAEP framework. Panelists were instructed to record alignment codes for all 42 items in their item booklets, and then to log in to the WAT and enter their codes electronically. Recording codes in item booklets was done to 1) minimize potential technical problems that might result from panelists being logged out of the WAT system during data entry, 2) create a hard-copy backup of all alignment codes in the event of electronic data loss, and 3) facilitate re-entry of DOK levels for these 42 items when they were mapped to the ACCUPLACER framework later in the week, by keeping a hard-copy record of each item's DOK level.

When their respective panelists completed mapping NAEP items to the NAEP framework, each facilitator reviewed her/his panelists' codes to ensure consistency of calibration and identify discrepancies in coding (i.e., items assigned to three different DOK levels or to two non-contiguous DOK levels, and/or items not assigned by more than half of the panelists to the same objective). Discrepant items were then adjudicated within each panel, with the explicit instruction that panelists were not required to reach consensus, and panelists entered their changes to their codes if necessary to reflect any changes in their coding judgments. This discussion of items with discrepant codes was done to determine whether differences were based on a misinterpretation or systematic difference in application of the protocol, were related to specific issues with an item or standard, or were random differences among panelists.

Panelists took a break after discussing and possibly changing their codes, during which time facilitators and project staff began preparing for cross-panel adjudication (the process of ensuring in real time that the panels were functioning as replicate panels). The first steps of this process were for WestEd staff to run the WAT overall results report and prepare the cross-panel adjudication workbook for review and discussion. The facilitators then met separately with WestEd project staff and, usually, the COR, to compare the results of the two groups for discrepancies as outlined in the design document. The facilitators and WestEd project staff reviewed the four alignment criteria: categorical concurrence (reviewing average numbers of items assigned to each objective), depth-of-knowledge consistency (reviewing average

[7] The project director collaborated with the two mathematics facilitators to select a representative range of sample items from the bank of released NAEP items (National Center for Educational Statistics, 2009) and the bank of released ACCUPLACER items (The College Board, 2007). The facilitators then independently coded and reached agreement on DOK levels and alignment to objectives for each item prior to the commencement of this study.
[8] The sample items, representing a range of DOK levels and objective alignments, were selected by the facilitators to ensure that both panels were introduced to a range of potential coding issues.

percentages of items at, below, and above the DOK level of the assigned objective), range-of-knowledge correspondence (reviewing the percentages of objectives with at least one aligned item), and balance of representation (reviewing index values). Per the design document, discrepancies of greater than five mean hits (categorical concurrence) or five percentage points (depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation), as well as balance of representation index values lower than .7, were investigated to determine whether the differences between panels were systematic or random. As directed by the design document, the facilitators first attempted to resolve areas of discrepancies by discussing observations and panelist opinions raised during the coding process that might have been related to the difference in results. Next, facilitators used the WAT reports to identify specific items that were coded differently by each panel, keeping in mind that panel results are an average across all eight panelists. When relevant items were identified, the facilitators discussed the items and determined whether the difference in coding was reasonable (i.e., not an error), and whether it was random or the result of a systematic difference in interpretation. Facilitators then reported back the outcomes of the cross-panel adjudication (i.e., areas of discrepancy, if any, and whether those discrepancies were systematic or random) to their respective panels, including raising specific items for discussion if necessary. Then, panelists were given the opportunity to change alignment codes if necessary to reflect any changes in their coding judgments. WestEd staff used these final alignment codes in the analysis. Areas of adjudication are discussed in the sub-study results.

Review the ACCUPLACER Framework

The design document developed to guide this project's pilot and operational studies calls for all coding to the NAEP framework to be completed before assigning DOK levels to ACCUPLACER objectives. However, following the pilot study, WestEd and the Governing Board, in consultation with Dr. Webb, determined that DOK levels should be assigned to each framework and that within-framework coding (i.e., mapping NAEP items to the NAEP framework, and mapping ACCUPLACER items to the ACCUPLACER framework) should occur before cross-framework coding (i.e., mapping NAEP items to the ACCUPLACER framework, and mapping ACCUPLACER items to the NAEP framework) occurred. This modification to the design was intended to allow panelists to code each assessment to its own framework before being exposed to the items through cross-framework coding. Since the ACCUPLACER framework did not contain sufficient detail to be coded for DOK, the next step in this alignment workshop's alignment process was for the panels to review the ACCUPLACER objectives to become familiar with the content prior to coding. The facilitators met with the combined mathematics panels to discuss the ACCUPLACER objectives, as well as preliminary decision rules related to the coding of the objectives. In order to proceed with the study using the WAT system, it was necessary to enter a DOK level for each objective. Therefore, facilitators entered "2" as the default level, although it was understood by all that this was not an actual DOK rating for the objectives.

Map ACCUPLACER Items to the ACCUPLACER Framework

As with the mapping of NAEP items to the NAEP framework, a smaller item set was used in the mapping of ACCUPLACER items to the ACCUPLACER framework. To refresh panelists in the use of alignment criteria, at the beginning of this task, each facilitator led her/his panelists

through the coding of a limited sample of active ACCUPLACER items[9] from the item booklet to ensure calibration among panelists. Once calibration was reached, panelists began to independently map the remaining ACCUPLACER items to the ACCUPLACER framework, recording codes both in item booklets and in the WAT. As described earlier for NAEP-to-NAEP item alignment, coding discrepancies were adjudicated both within and between the two panels. Within-panel discussions focused on items coded at more than two DOK levels, items coded at non-adjacent DOK levels, and items for which there was no majority of objective codes. Items were discussed, but consensus was not required. Cross-panel adjudication focused on alignment criteria for which there was a discrepancy between panels of greater than five percentage points. Again, consensus was not required, but any issues were communicated to panelists, who had the option of changing their codes. These final alignment codes were used by WestEd staff to determine if the results from the two panels were, indeed, random and not the result of systematic differences in the application of the protocol or the framework or misinterpretations of the protocol, framework, or items.

Map NAEP Items to the ACCUPLACER Framework

The procedures described earlier for mapping each assessment's items to its framework were used to map NAEP items to the ACCUPLACER framework, although for this alignment task the entire NAEP item pool was used. Because the first 42 NAEP items had been assigned DOK levels when being mapped to the NAEP framework, those assigned DOK levels were re-entered into the WAT for this task; thus, for the first 42 items, the task of mapping to the ACCUPLACER framework was limited to determining alignment to objectives. For all remaining NAEP items, within this task, DOK levels were assigned and alignment to objectives was determined.

Map ACCUPLACER Items to the NAEP Framework

The procedures described earlier were used to map ACCUPLACER items to the NAEP framework objectives. Because the first 45 ACCUPLACER items had been assigned DOK levels when being mapped to the ACCUPLACER framework, those DOK levels were re-entered into the WAT for this task; thus, the task of mapping the first 45 ACCUPLACER items to the NAEP framework was limited to determining alignment to objectives. For the remaining 60 items, this task included both assigning DOK levels and determining alignment to objectives.

Pacing and Schedule Adjustments

Throughout the week, the mathematics panelists worked beyond the planned adjournment time each day to complete their work. The time-intensive schedule was anticipated by WestEd and the COR because of the large number of test items and objectives in both mathematics assessments, information gathered in the pilot study, the experiences of the ACT WorkKeys study, and the time estimates in the study design document. Attempts to ameliorate this included the use of a sample pool for the within-framework sub-studies, as described earlier. WestEd staff monitored the schedule closely, and, in consultation with the COR and facilitators, adjusted the daily schedule as needed. Schedule adjustments were made based on a number of factors, including

[9] The sample items, representing a range of DOK levels and objective alignments, were selected by the facilitators to ensure that both panels were introduced to a range of potential coding issues.

the importance of keeping the panels synchronized in the tasks they were completing (one panel was not permitted to move ahead to a new coding task before the other had completed it, in case issues arose during cross-panel adjudication that would impact subsequent tasks). WestEd staff tried to allow sufficient break time before starting new tasks, so that panelists would be refreshed and ready to code. Due to the large number of items and objectives to be coded, however, it was necessary to ask panelists to stay past the scheduled end time and/or start earlier in the morning in order to complete the tasks. Panelists were cooperative and started early and stayed late as needed in order to finish each day's tasks. Additionally, some panelists requested assistance with the data entry of their handwritten codes, either from project staff or from other panelists who had already finished a task. All alignment tasks were completed by both panels by the end of the alignment workshop.

Decision Rules

During the framework analysis and item review conducted prior to the alignment workshop, facilitators developed a preliminary set of decision rules for use by panelists. Facilitators reviewed the preliminary decision rules with panelists and instructed panelists in their use prior to alignment coding, ensuring that panelists were comfortable with the decision rules. Throughout the alignment coding sessions, additional decision rules could be developed and existing decision rules modified if doing so was necessary to clarify potential ambiguities in assessments and assessment frameworks, thereby promoting consistency in coding both within and across panels; any additions and modifications were carefully considered by the facilitators and agreed to by both panels. The final list of decision rules used for this alignment workshop follows.

NAEP Mathematics Framework for Alignment: Decision Rules

1. The objectives within the "Algebra" standard will be interpreted as aligning primarily to items containing one or more variables and not items containing only numerical expressions.
2. The objectives within the "Number operations and properties" standard will be interpreted as aligning primarily to numerical items; however, consideration is also to be given to items containing one or more variables.
3. The primary intent of objectives containing wording similar to the following is to assess mathematics in [problem-solving situations (in either a real-world or mathematical context), as opposed to the performance of simple procedures or algorithms].[10]
    - 1.1.g, "Represent, interpret, or compare expressions or problem situations involving absolute values"
    - 1.3.f, "Solve application problems involving numbers, including rational and common irrationals"
    - 1.4.c, "Use proportions to solve problems (including rates of change)"

[10] Text in brackets refers to clarifications made for a more general audience.

4. Some objectives contain multiple parts separated by the word “and” (see 1.5.f below). The intent of the objective may or may not be to assess all parts. If an item addresses only one part of the objective, panelists are asked to look for an alternative primary code. If an alternative is not available, panelists are to note in the WAT that the item does not assess the entire objective.
    - 1.5.f, “Recognize properties of the number system (whole numbers, integers, rational numbers, real numbers, and complex numbers) and how they are related to each other, and identify examples of each type of number”
5. An objective that addresses expressions may also be aligned with an item containing an equation if symbolic manipulation across the equal sign is not required to answer the question.

ACCUPLACER Mathematics Framework for Alignment: Decision Rules

1. When coding an item, panelists will consider the intended audience for objectives. For example, a likely place for a decision of this sort is among the algebra objectives shared by both the Elementary Algebra test and the College Level Math test. If the item assesses algebra at a basic level, the item should be aligned to an “Elementary algebra” objective. If the item assesses algebra at an advanced level, the item should be aligned to a “College level math” objective.
2. As with the NAEP objectives, the primary intent of ACCUPLACER objectives containing wording similar to the following is to assess mathematics in [problem-solving situations (in either a real-world or mathematical context), as opposed to the performance of simple procedures or algorithms].
    - A.3.a, “Rate problems including ratio and proportion”
    - A.3.b, “Percent problems”
    - A.3.d, “Measurement problems”

Participants

WestEd Staff and Respective Roles

The project management team on-site for this study comprised Mr. Peter Worth (project director), Dr. Stanley Rabinowitz (principal investigator), Dr. Jennae Bulat (project coordinator), Mr. Greg Hill, Jr. (coordinator), and Ms. Jennifer Verrier (administrative assistant).

As project director, Mr. Worth executed day-to-day project management, including managing the schedule and budget, overseeing project staff, and directing all communication with the Governing Board COR.

Working closely with Mr. Worth, Dr. Rabinowitz provided intellectual leadership, including spearheading up-front planning of the overall study; overseeing development of protocols, procedures, and materials; and reviewing all reports.

Dr. Bulat worked with Mr. Worth to oversee day-to-day work, coordinate and support the work of the alignment panels, supervise arrangements for travel and facilities, and contribute to this comprehensive report.

Mr. Hill provided logistical and technical support to project management, coordinating the production of study materials to management specifications. He also developed technical resources to support reporting processes and data analysis.

Ms. Verrier, a WestEd staff member working out of WestEd's Washington, DC, office, provided on-site logistical and technical support to project management, assisting with study material management, overall logistical management, facility coordination, and data entry.

Facilitators and Facilitator Qualifications

The two facilitators recruited for this study played key roles on the project team, developing and/or vetting all materials to be used by the panels, training both sets of panelists, ensuring calibration with the Webb content and complexity evaluation criteria, and working closely with and training other WestEd staff to ensure consistency and dependability in the completion of project tasks.

Mr. Michael Brown served as lead mathematics facilitator for this study, conducting the comparative analysis of the NAEP and ACCUPLACER frameworks, leading one of the two study panels, working with the second mathematics facilitator to reach agreement (where necessary) and resolve differences in interpretation across panels throughout the study, and playing a key role in writing and reviewing the results section of this report. Working for WestEd, Mr. Brown has served as a consultant on alignment studies for multiple states and consortia. Additionally, he has developed state assessments for multiple states. His previous work involved teaching mathematics at grades 6 through 12 and community college for 13 years. He has also served as an assessment specialist for a testing company and conducted presentations on mathematics content and pedagogy. Mr. Brown holds a BA in general and comparative studies with an emphasis in biology and liberal arts from the University of Texas, Austin, and a MEd in secondary education from Southwest Texas State University.

Ms. Linda McQuillen served as the second mathematics facilitator for this study, leading one of the two study panels and working with the lead facilitator to reach agreement (where necessary) and resolve differences in interpretation across panels throughout the study. Ms. McQuillen has worked with Dr. Norman Webb as a reviewer on 18 mathematics and special education alignment studies in eight states and the country of Qatar. She has also conducted a professional development seminar for Marion County, West Virginia, on depth of knowledge and the alignment process. For the past four years, Ms. McQuillen has worked at the University of Wisconsin, Madison, School of Education, as an associate lecturer in the Department of Curriculum and Instruction. She has taught for over 40 years, specializing in the areas of mathematics and special education. Ms. McQuillen received a BS in secondary education from Northern State College, and a MS in exceptional education from the University of Wisconsin, Milwaukee.

Panel Criteria for Recruitment and Panelist Qualifications

A total of sixteen panelists, eight for each of the two replicate panels, were recruited for participation in the operational alignment workshop. The following criteria were used to recruit panelists:

- Deep knowledge of the subject matter, as exemplified by relevant academic degrees and a range of training and experiences; at least 5–7 years direct experience with high school and lower-level postsecondary students in the content area; and/or experience in reviewing, analyzing, and/or developing curricula, standards, and/or assessments in the content area.
- Experience in reviewing, analyzing, and developing curricula, standards, and assessments, especially at the secondary and postsecondary levels.

In order to ensure that the panelists did not hold biases toward any of the assessments included in the study, panelists with substantial involvement in the development of either NAEP or ACCUPLACER were disqualified from participation in the alignment workshop. In addition, WestEd sought panelists who would represent a range of knowledge of each assessment on each panel.

As agreed upon by the Governing Board, nominations were solicited and panelists were recruited from the following sources:

- Referrals from the 2009 NAEP Mathematics Specifications Work Group, the 2005 NAEP Mathematics Project Steering Committee, and the 2005 NAEP Mathematics Project Planning Committee, as identified in the 2009 NAEP mathematics framework (National Assessment Governing Board, 2008).
- WestEd's immediate network of state and district educators, administrators, coordinators, and other content area experts from across the country who have worked with WestEd on alignment, assessment, and standards review projects.
- National education professional organizations, such as the National Council of Teachers of Mathematics, the National Council of Supervisors of Mathematics, and the Association of State Supervisors of Mathematics.
- Departments of mathematics and schools of education from top-ranked colleges and universities across the country.[11]

[11] Regional and national colleges and universities were targeted as resources for nominators and/or potential panelists. Institutions were selected based on rank and expertise as rated by *U.S. News and World Report* (2010) (e.g., top fifty nationally recognized PhD-granting institutions and top regional master's-degree-granting institutions). Department heads from top-tier national and regional institutions were contacted to solicit referrals and to recruit as potential candidates.

Panels were structured to achieve the desired balance of secondary and postsecondary professional experience (including both current and prior experience) among panelists:

- On the first panel, 50% of panelists (4 of 8) reported experience in both secondary and postsecondary mathematics education; 25% (2 of 8) had secondary teaching experience only; and 25% (2 of 8) had postsecondary teaching experience only.
- On the second panel, 63% of panelists (5 of 8) reported experience in both secondary and postsecondary mathematics education; 38% (3 of 8) had secondary teaching experience only.

The composition of panels was balanced according to background expertise and experience with the NAEP and ACCUPLACER assessments. Every attempt was made to balance each panel by geographic representation, race, ethnicity, and gender, although panelist availability limited the results of these attempts. The distribution of gender was comparable across the panels, with four women on each of the eight-member panels, as was representation of advanced degrees, with six and five doctoral degrees represented on Panels 1 and 2 respectively. Panelists represented a range of geographic areas, including the Northeast (Pennsylvania), the South (Florida, Kentucky, Louisiana, Mississippi), the Midwest (Indiana, Michigan, Missouri), and the West (California, New Mexico, Texas). WestEd was unable to achieve race/ethnicity diversity, however. All panelists but three identified themselves as White/Caucasian/of European descent; two panelists identified themselves as Asian/Caucasian; and one panelist identified herself as Black/African-American. A list of panelists organized by panel follows.

Mathematics Panel 1

[redacted]

[redacted]

Mathematics Panel 2

[redacted]

Preparation, Materials, and Logistics

Facilitator Training

Prior to the NAEP–SAT alignment workshop held in March 2010, an initial facilitator training was held to introduce the objectives of the project as a whole and the alignment criteria and methodology to be used across all alignment workshops. The facilitators were asked to review the study design document and the *Web Alignment Tool* (*WAT) Training Manual* (Webb, 2005) in preparation for that training. Prior to the NAEP–ACCUPLACER mathematics alignment workshop, a follow-up facilitator training session was held. Facilitators were asked to re-review both documents prior to this training, specifically to enhance their familiarity with use of the WAT system. The facilitators had in-depth knowledge of the two frameworks. The lead analyst had analyzed the two assessment frameworks for the NAEP–ACCUPLACER interim report. They were also asked to review the NAEP and ACCUPLACER frameworks and both sets of assessment items in order to identify potential coding challenges and draft decision rules. Both facilitators selected for this study are well versed in alignment methodologies. They had participated in the NAEP–ACCUPLACER reading pilot study and thus had been previously trained in the objectives of this project and the alignment criteria to be used across all operational studies. WestEd, therefore, emphasized the following in the follow-up training:

- Review of alignment workshop objectives and design overview
- Agenda review
- NAEP and ACCUPLACER assessment overview and discussion of issues
- NAEP and ACCUPLACER framework overview and discussion of issues
- Discussion of NAEP and ACCUPLACER decision rules
- Facilitator roles and responsibilities (e.g., security protocols)
- WAT system use

Materials from both facilitator training sessions are included in Appendix G.

Pre-Workshop Facilitator and Panelist Materials

In preparation for the NAEP–ACCUPLACER mathematics study, the study's lead facilitator developed the comparative analysis to document the similarities and differences between the NAEP and ACCUPLACER mathematics frameworks. Prior to this alignment workshop, the facilitators reviewed the NAEP and ACCUPLACER frameworks and discussed the results of the comparative analysis. The facilitators and WestEd's project management identified issues that might impact alignment coding, and they developed preliminary decision rules to guide panelists. Approximately two weeks prior to the alignment workshop, both facilitators received NAEP and ACCUPLACER items to code in advance of the alignment workshop, again to identify issues to address with panelists.

Also approximately two weeks prior to the alignment workshop, panelists were sent a draft agenda overview, NCES and College Board confidentiality agreements, the *Mathematics Framework for the 2009 National Assessment of Educational Progress* (National Assessment Governing Board, 2008), a College Board *ACCUPLACER: Revealing Potential. Expanding Opportunity* brochure (College Board, 2009), and additional background information on the

ACCUPLACER printed from the College Board website (College Board, 2010). In an accompanying cover letter, panelists were asked to review the documents prior to the start of the alignment workshop to ensure that they were familiar with the content of the assessments.

Facilitator and Panelist Binder Materials

Once on-site, each facilitator and panelist received a binder that included both logistics documentation (i.e., an agenda, NCES and College Board confidentiality agreements, travel and other expense reimbursement forms, and a list of panelist names) and training materials (i.e., a copy of the training PowerPoint presentation, alignment coding information, WAT training materials, sample items for alignment training, and a blank assessment coding form). The facilitator binders also contained an excerpt of depth of knowledge coding procedures from the *WAT Training Manual* (Webb, 2005) and a facilitator alignment process guide developed by WestEd. Abbreviated versions of the panelist binder (excluding expense reimbursement forms) were made available for observers to use on a daily basis. A copy of the alignment workshop's daily agenda is provided in Appendix D. Copies of facilitator training materials are provided in Appendix G.

Panelist Training Materials

Panelist training for assigning DOK levels to objectives occurred on the first morning of the alignment workshop. Panelist training for assigning DOK levels of items and for coding items to objectives occurred on the second morning of the workshop. In addition, facilitators reviewed the alignment criteria at the beginning of each alignment session and provided refresher training as needed. A combined (reading and mathematics) panel training session introduced the purpose of the overall study and the NAEP and ACCUPLACER assessments; it also provided an overview of the alignment process, definitions of alignment criteria, and use of the WAT (copies of panelist training materials are provided in Appendix E). Following this introduction and overview, the two reading panels relocated to a separate room, and the combined mathematics panels received training on assigning DOK levels to objectives, using practice objectives drawn from the *WAT Training Manual* (Webb, 2005). Additional training on assigning DOK levels to items and assigning items to objectives was subsequently provided, using sample items drawn from the *NAEP Sample Questions, Grade 12, 2009* (National Center for Educational Statistics, 2009), from the *NAEP Sample Questions, Grade 12, 2005* (National Center for Educational Statistics, 2005), and from the *ACCUPLACER Sample Questions for Students* (The College Board, 2007). These sample items were selected by the mathematics facilitators to represent a range of item types, DOK levels, and objective alignments, and are included in Appendix E.

On-Site Security of Materials

WestEd secured frameworks, anchor papers, and all other secure materials in locked rooms when not under direct WestEd staff supervision. Otherwise, all meeting rooms containing secure materials were constantly attended to by WestEd staff or content facilitators. WestEd developed a security protocol to document and enforce the level of test material security required by this study, including the areas listed below:

- Shipping and receiving of materials to the Westin
- Meeting room security

- Panelist, facilitator, and observer confidentiality agreements
- Secure management of test materials on-site
- Secure management of WAT reports on-site

A copy of this security protocol and the secure materials tracking sheets are provided in Appendix H.

Item Booklets, Framework Documents, and Anchor Papers

For the NAEP item booklets, the 164 items were organized in block and item order, numbered sequentially within each block, except that the 42 items identified for coding to the NAEP framework were listed first, also in block and item order. Each item was presented on a separate page, with grade, block code, NAEP identification and WestEd sequence numbers, item type, and answer key indicated at the top of the page. On each page, WestEd demarcated an area in which that item's DOK rating, NAEP alignment code, and ACCUPLACER alignment code were to be recorded.

For the ACCUPLACER item booklets, the three core mathematics tests were combined, with items numbered sequentially. The first 45 items were the 15 common items from the three tests, in the order in which they appear on their respective forms. These were followed by the remaining 20 variable items from the three tests, also in the order in which they appear on their respective forms, for a total of 105 items. Items were not labeled according to their core test, but rather as one item booklet. Each item was presented on a separate page, with item sequence number, College Board item identification number, and answer key indicated at the top of the page. On each page, WestEd demarcated an area in which that item's DOK rating, NAEP alignment code, and ACCUPLACER alignment code were to be recorded.

WestEd staff made available individual copies of the NAEP and ACCUPLACER frameworks, which facilitators and panelists checked out on a daily basis. These versions of the framework documents provided space for the DOK rating of each objective to be noted.

The NAEP item booklets included detailed scoring information for each constructed-response item. In addition, WestEd staff provided a set of NAEP anchor papers (sample student responses at each score point for each constructed-response item) to each panel for use in determining the intended level of student response on constructed-response items. Panelists were encouraged to use the anchor papers as needed to help determine the intent of any given constructed-response item, although they were not required to do so. Facilitators reported that, in practice, panelists found the items and scoring information sufficient to determine item DOK and alignment to objective, and that the anchor papers were rarely consulted for this purpose.

All secure documents, including item booklets and frameworks, were color-coded and visibly marked as being secure.

Questionnaires and Final Debrief

In addition to the item alignment ratings captured in the WAT, panelists were surveyed throughout the five-day alignment workshop to 1) determine their judgment of alignment for each alignment activity (e.g., NAEP assessment to NAEP framework) in lieu of the similar

debrief surveys that exist within the WAT itself (debrief questionnaires), and 2) evaluate the effectiveness of the overall alignment process and alignment workshop logistics (e.g., needs for additional information, adequacy of the facility) (process questionnaires). Both debrief and process questionnaires are included in Appendix F. Process questionnaires are discussed in Section IV of this report.

A full-group debrief and discussion at the end of the week provided an opportunity to evaluate the overall alignment process, evidence generated, criteria applied, and holistic conclusions regarding alignment of the assessments; generate recommendations regarding alignment and appropriate use of evidence; and evaluate panelists' understanding of procedures.

Debrief Questionnaires

- A debrief questionnaire was administered immediately following *each* coding session's alignment of a set of items to a framework in order to solicit feedback regarding that alignment coding session. These debrief questionnaires solicited specific feedback regarding the coding of each set of assessment items to each framework as a supplement to the alignment codes captured within the WAT system. In a typical WAT-based alignment study, these questionnaires would be administered online as part of the WAT system; however, as this study's design called for a modified set of questions, debrief questionnaires were administered in a paper format, and panelists were instructed to complete the paper versions instead of the questionnaires presented in the WAT system. Within the WAT, panelists were required to respond to one of the WAT debrief questionnaire questions in order to complete their coding sessions; therefore, panelists were instructed to respond online to WAT question D, indicating their judgment of overall alignment, as well as answering the same question on their paper-based debrief questionnaire.
- An end-of-framework questionnaire was administered at the completion of all coding to the NAEP and ACCUPLACER frameworks. These questionnaires solicited feedback regarding similarities and differences between the two assessments relative to the respective framework and regarding the functionality of the framework organization.

Process Questionnaires

- A training questionnaire was administered following panelists' training on the first day of the alignment workshop to solicit feedback on the training's effectiveness and to identify areas in which more information might be needed. This questionnaire was administered via an online survey system (SurveyMonkey).[12]
- An evaluation-of-process questionnaire was administered at or near the end of each of the second, third, and fourth days of the alignment workshop. These questionnaires were used to monitor panelists' understanding of the process, and to solicit questions, concerns, and other feedback from panelists regarding that day's activities. These questionnaires were administered via the online SurveyMonkey system.
- An end-of-workshop questionnaire was administered at the end of the week to solicit feedback regarding the meeting logistics (e.g., meeting rooms, food, equipment), the alignment process (e.g., training, materials, adjudication procedures, use of the WAT), and differences observed between the two assessments. To protect any secure comments

[12] http://www.surveymonkey.com

that might have been made on this questionnaire, this questionnaire was administered in a paper format.

These questionnaires captured important information about both alignment and process. WestEd staff evaluated the results of the process questionnaires at the end of each day in order to monitor panelists' perceptions of and comfort with the alignment process and to identify areas of concern and/or needs for additional training; these results are summarized in Section IV of this report. Full responses to the process questionnaires are in Appendix I. Debrief questionnaires capture important qualitative information regarding alignment coding, which was used to help inform conclusions about the alignment between each framework/assessment pair. Full responses to the debrief questionnaires are in Appendices J–M.

Final Debrief
As the final task of the week, the combined panels convened with the two facilitators, WestEd staff, and Governing Board observers to discuss how the process captured the content similarities/differences between the assessments, to what degree the two assessments aligned, and, considering the items in each assessment, how the assessments were the same and/or differed. This final debrief session also provided an opportunity for panelists to express any thoughts, concerns, or questions that remained regarding the assessments, objectives of the overall study, and projected use of study results.

WAT System

As indicated earlier, the WAT system was used to record alignment ratings, analyze data, and generate reports for this alignment workshop. Prior to the commencement of the alignment activities, WestEd staff set up each panel as a group within the WAT, entered into the WAT the NAEP and ACCUPLACER items (i.e., assigned item numbers and item weights) and frameworks (i.e., standard, goal, and objective labels, organized into the WAT three-level hierarchy), and created the four requisite WAT studies for each group:

- NAEP (short version) items to NAEP framework
- ACCUPLACER (short version) items to ACCUPLACER framework
- NAEP items to ACCUPLACER framework
- ACCUPLACER items to NAEP framework

Facilities

This alignment workshop was held at the Westin Grand hotel in Washington, DC. The hotel was contracted to provide all guest and meeting rooms, technical support, ancillary technical equipment (e.g., hubs, power strips), and food and beverage catering. A separate vendor was contracted to provide laptop computers for facilitators and panelists, printers, and projector screens. All other equipment was provided by WestEd.

Mathematics panels used two hotel meeting rooms throughout the alignment workshop. Because this alignment workshop ran concurrently with the NAEP–ACCUPLACER alignment workshop in reading, a meeting room large enough to accommodate all reading and mathematics panelists was used for whole-group training and adjudication sessions. This large room was also used for combined mathematics panel training and discussion, and as the coding room for one

mathematics panel. A smaller room was used by the other mathematics panel for coding sessions.

Each room was equipped with a printer and nine working stations (eight panelist stations and one facilitator station), each one comprising a laptop, mouse, high-speed Internet connection, and working space. Each room also supported the use of an LCD projector, as needed or desired by the facilitator. When housing secure materials, each room was locked when not supervised by a facilitator or a WestEd staff member. All rooms were locked at the end of each working day. Space was provided at the back of each meeting room to accommodate approved observers (i.e., Governing Board staff, College Board staff, and a technical advisor), who were free to observe panels at their discretion.

Pilot Study: Lessons Learned

As stipulated by the Governing Board, a preliminary study was conducted to pilot test the methodology and logistics proposed for the four operational alignment studies. It was agreed by WestEd and the Governing Board that the pilot study would focus on the grade 12 NAEP and ACCUPLACER assessments in reading. This content area and assessment pairing was selected in order to address the complexities associated with computer-adaptive assessments (e.g., identifying an appropriate item pool) and the complexities associated with the content area of reading (e.g., reading genres, reading purpose, and the role of passages). In doing so, the most complex aspects of the methodology—including coding procedures, data analyses, training and alignment protocols, materials, and logistics—would be evaluated. The pilot study was conducted from December 14–18, 2009, in Washington, DC. The size of each panel was limited to four for the purposes of the pilot study, although all other aspects of the study matched the design and implementation of the operational studies as closely as possible. Although the focus of the pilot was reading, the two mathematics facilitators participated in the pilot study as observers, in order to see firsthand the study design being implemented. By observing the first two days of the pilot, the mathematics facilitators were able to see the training, coding, and initial review of cross-panel results for one study. A full accounting of that pilot study can be found in WestEd's Pilot Study Report, submitted to the Governing Board on March 19, 2010, and a summary of the recommendations based on the pilot study follows. Although some of the recommendations are specific to the content area of reading, they are included here to preserve the completeness of the list, and because of the potential for lessons to be transferred to mathematics.

Sequence of Study Steps
Modify the coding order to code DOK levels of both sets of frameworks prior to the coding of their respective sets of items. This is intended to make the process more comparable for the two frameworks and help to eliminate any potential related bias or influence over the DOK coding process caused by having analyzed Pexam (the generic term used for the performance exams to which NAEP is compared) items prior to analyzing the Pexam framework.

Within-Panel Adjudication
Facilitators may share their own alignment interpretations to foster group discussions and help clarify understandings and interpretations, but care should be taken to ensure the facilitator's interpretation does not dominate or overly influence that of the panelists.

Preserve the table space of the "classroom" setup and instruct panelists to face one another during discussion.

Cross-Panel Adjudication
Refine and use WestEd's Excel workbook tool to present and compare the results of the two replicate panels in order to inform cross-panel adjudication discussions.

Questionnaires
To minimize panelist fatigue, limit the number of questionnaires administered to panelists by consolidating training and process evaluation questionnaires as much as possible.

Administer training and process questionnaires, which do not contain or solicit sensitive information, via an online survey engine for greater panelist convenience.

Frameworks
Refine the organization and presentation of the NAEP reading framework document used for coding (e.g., consolidate redundant objectives, revise wording of objectives) to reduce ambiguity and/or redundancy.

Identify and provide additional information, if available, to elaborate on the ACCUPLACER frameworks used for coding.[13]

Facilitator Training
Provide facilitators with assessment frameworks and sample items for review at least two weeks in advance of the study. As facilitators code sample assessment items to the frameworks, they will identify any preliminary decision rules and determine where coding and adjudication discrepancies and areas of potential confusion might exist prior to the study.

Refine facilitator training to include additional training on the WAT system, tailored specifically for this study, and the use of the WestEd Excel workbook tool as well as the logistics of the methodology.

Panelist Training
Provide frameworks and other preparatory materials to panelists at least two weeks prior to the study as mandatory reading material for the session.

Refine panelist training to address and/or emphasize the areas identified in the Pilot Study Report as needing clarification or specifications: alignment criteria, including examples in areas such as clarification of the definition(s) of a match, especially to multiple objectives; the operational difference among primary/secondary/uncodable item codes; the differentiation between complexity and difficulty; the need to consider knowledge and skills rather than the ability of an individual student; and the distinction between cognitive targets and DOK levels.

Provide more training on the use of the WAT system (e.g., the interface, screens for each step in the process, and how to code and track common items).

[13] This recommendation proved necessary for the SAT reading and mathematics and ACCUPLACER mathematics frameworks as well.

Remind panelists to read the reading passages each time they are coding their respective items to maximize consistency across coding.

Materials
Revise the ACCUPLACER objective numbering scheme to avoid confusion with DOK ratings.

Where possible, have materials available in larger print.

Schedule
Review and refine the agendas, including break and meal times, after a thorough review of the materials for the operational studies for each content area.

Equipment/Technology
Should technical difficulties arise with the WAT reporting, facilitators will implement the necessary steps of printing the raw data codes for each panelist and ensuring accurate data re-entry.

Analysis
Clarify and document the process for averaging or aggregating results across the two panels outside the WAT.

Combine the ACCUPLACER forms into one item pool for the operational studies, including the common items only once, in their first position, and assign them a double weighting to retain the accuracy of the proportions. Make cross-assessment comparisons at the item pool level.[14]

All recommendations were implemented.

[14] This recommendation is relevant to reading only. For ACCUPLACER mathematics, the 15 common items from each of the three core tests (a total of 45) were combined as one form and used for the within-framework analysis. All items from one form (35) of each of the three core tests (a total of 105) were combined and used for the cross-framework analysis. Because there was no overlap of items across the core tests, no weighting was required.

III. Alignment Results

This section presents the results of the NAEP-to-ACCUPLACER alignment study. The section begins by reporting the interrater agreement within panels. Then, the DOK of the NAEP framework and NAEP and ACCUPLACER assessment items are discussed. Finally, the results of the four sub-studies are presented.

Reliability and Interrater Agreement

The degree to which panelists within a panel assigned the same codes to the items is presented with four measures of interrater agreement. Consensus of item codes among panel members was neither a requirement nor a goal of this study. However, as described in Section II of this report, it was important that panelists discuss items for which there was a wide discrepancy of DOK levels (i.e., items assigned to more than one level or to non-adjacent levels) or matches to objective (i.e., items with no majority agreement of ratings) among panelists, to determine whether differences were based on a misinterpretation or systematic difference in application of the protocol, were related to specific issues with an item or standard, or were random differences among panelists.

Table 2 shows the interrater agreement for each panel for each sub-study, as reported by the WAT (full WAT reports by sub-study are provided in Appendices J–M). Interrater agreement is provided to indicate the degree of reliability of both DOK ratings and the coding of objectives and standards to items. For DOK ratings, interrater agreement is determined through the calculation of intraclass correlation and pairwise comparison statistics. As described by the *WAT Training Manual*, the intraclass correlation statistic "measures the percent of variance in the data due to the differences between the items rather than the differences between the reviewers" (Webb, 2005, p. 115). Values are considered in the highest range in which they fall; values greater than 0.7 reflect adequate agreement, while values greater than 0.8 reflect good agreement. Because low variance among the items can make the intraclass correlation statistic misleading, the WAT also provides pairwise comparison values (p. 115). The WAT calculates pairwise comparison for DOK by comparing the ratings assigned by each possible pair of panelists in a panel, dividing the number of agreeing pairs by the total number of pairs, and then finding the average agreement across all items on a test. Values of 0.7 or higher reflect good agreement, values of 0.6 or higher reflect reasonable agreement, and values lower than 0.5 reflect poor agreement (p. 116).

Pairwise comparison statistics are also calculated to show the interrater agreement for panelists' judgments of alignment of items to objectives in the frameworks. Interrater reliability of these judgments is reported at the more specific objective level (i.e., the degree to which panelists reached the same judgment of the objective[s] tested by an item) and the more general standard level (i.e., the degree to which panelists reached the same judgment of the standard containing the objective[s] tested by an item). Objective and standard pairwise comparison are calculated as follows: for each pair of reviewers, "find the reviewer who coded the greater number of objectives to this item, and call this number *n*. Now take the number of entries the two reviewers agree on and divide this by *n*. This is the *agreement* between the two reviewers. Perform this calculation for all possible pairs of reviewers, and take the sum of the agreements. Then divide this sum by the total number of pairs of reviewers. This is the *pairwise agreement* value for the given assessment item . . . The pairwise agreement for objectives is averaged over all the

assessment items to give the *pairwise agreement for objectives* statistic for the alignment study as a whole" (Webb, 2005, p. 115). It is typical that objective pairwise comparison values are lower than those for standard pairwise comparison, because objectives tend to be more specific applications of a broader topic defined in a standard.

Table 2. Interrater Agreement of Panels by Sub-Study

Sub-Study	Panel 1	Panel 2
Sub-Study 1: NAEP to NAEP	*DOK* Intraclass Correlation: 0.9201 Pairwise Comparison: 0.6862 *Objective, Standard* Objective Pairwise Comparison: 0.4798 Standard Pairwise Comparison: 0.8673	*DOK* Intraclass Correlation: 0.9406 Pairwise Comparison: 0.6905 *Objective, Standard* Objective Pairwise Comparison: 0.6943 Standard Pairwise Comparison: 0.9838
Sub-Study 2: ACCUPLACER to NAEP	*DOK* Intraclass Correlation: 0.8493 Pairwise Comparison: 0.6765 *Objective, Standard* Objective Pairwise Comparison: 0.6509 Standard Pairwise Comparison: 0.9454	*DOK* Intraclass Correlation: 0.9209 Pairwise Comparison: 0.7789 *Objective, Standard* Objective Pairwise Comparison: 0.752 Standard Pairwise Comparison: 0.9352
Sub-Study 3: ACCUPLACER to ACCUPLACER	*DOK* Intraclass Correlation: 0.8934 Pairwise Comparison: 0.7302 *Objective, Standard* Objective Pairwise Comparison: 0.7009 Standard Pairwise Comparison: 0.9294	*DOK* Intraclass Correlation: 0.9396 Pairwise Comparison: 0.8048 *Objective, Standard* Objective Pairwise Comparison: 0.77 Standard Pairwise Comparison: 0.9337
Sub-Study 4: NAEP to ACCUPLACER	*DOK* Intraclass Correlation: 0.9018 Pairwise Comparison: 0.7001 *Objective, Standard* Objective Pairwise Comparison: 0.4082 Standard Pairwise Comparison: 0.7344	*DOK* Intraclass Correlation: 0.9167 Pairwise Comparison: 0.7034 *Objective, Standard* Objective Pairwise Comparison: 0.6614 Standard Pairwise Comparison: 0.8238

Looking across panels, Table 2 shows that interrater agreement (within-panel) values for each panel were comparable for DOK and for alignment for most studies. Interrater agreement for DOK (intraclass correlation and pairwise comparison) was good for all sub-studies for both panels. Likewise, standard pairwise comparison values were good for all sub-studies for both panels. For match to objective, objective pairwise comparison values were good or reasonable for all sub-studies except for Sub-Studies 1 and 4 in Panel 1. For Sub-Study 1 (NAEP to NAEP), Panel 1 had an objective pairwise comparison value below the "reasonable" range and on the threshold of the "poor" range, with a value of 0.4798. For Sub-Study 4 (NAEP to ACCUPLACER), the objective pairwise comparison value was considered "poor," at 0.4082.

Lower objective pairwise comparison values can result from overlapping or unclear objectives within or across standards, as well as from items being coded to multiple objectives. Indeed, facilitators and panelists found that there were overlapping objectives in the NAEP and

ACCUPLACER frameworks and that test items were sufficiently complex to be reasonably aligned to a number of different objectives, often within the same goal. However, it is also apparent that, in general, Panel 1 had lower objective pairwise comparison values than Panel 2. That is, following within-panel discussions, Panel 2 tended to come to greater agreement about alignment to specific objectives than did Panel 1. For example, in Sub-Study 4, several items found by most or all panelists in Panel 2 to be uncodable were mapped by at least some panelists in Panel 1 to one or more objectives (including broad objectives such as A.3.g, "Other non-routine problems"). Although both panels received the same training and instructions and applied the same criteria, facilitation style may have contributed to the tendency for Panel 2 to reach majority agreement more than in Panel 1. The study required that a large number of items be coded during tight time constraints. This had an impact on the number of items that could be discussed and the amount and type of discussion for each, and may have been a factor in the cross-panel difference. It is also possible that individual panelists or items may have contributed to these differences. Nonetheless, for all sub-studies, pairwise agreement at the standard level was considered "good" according to the Webb alignment criteria definitions. Overall, the interrater agreement levels warrant confidence in the reliability of each panel's findings and the overall conclusions of the study.

As described in Section II of this report, the degree of cross-panel agreement attained was monitored throughout the study, as stipulated in the study design document. Where specific points of discrepancy and adjudication occurred, these are discussed in the context of each sub-study.

DOK Levels of the NAEP Framework

Panelists assigned DOK levels to each objective in the NAEP framework. The within-panel DOK ratings were then compared across panels, and the two facilitators reached consensus on the final DOK ratings for each objective, discussing them with the combined mathematics panels as appropriate. Consensus DOK values for the NAEP framework are shown in Table 3. DOK ratings were assigned to the 130 specific objectives. These ratings are reported at the goal and standard level in the table. DOK ratings for each objective can be found in Appendix C. As explained in Section II, the ACCUPLACER framework was classified in categories that were not further defined in terms of the cognitive level at which the knowledge and skills would be assessed and, therefore, could not be coded for DOK.

Table 3. DOK Findings for the NAEP Mathematics Framework

NAEP Framework	# of Objectives	# and % of Obj. at DOK 1	# and % of Obj. at DOK 2	# and % of Obj. at DOK 3	Average DOK
1.1	4	2 (50%)	2 (50%)	-	1.5
1.2	3	1 (33%)	1 (33%)	1 (33%)	2
1.3	5	3 (60%)	2 (40%)	-	1.4
1.4	2	-	2 (100%)	-	2
1.5	4	2 (50%)	2 (50%)	-	1.5
1.6	2	-	-	2 (100%)	3

NAEP Framework	# of Objectives	# and % of Obj. at DOK 1	# and % of Obj. at DOK 2	# and % of Obj. at DOK 3	Average DOK
1 overall	**20**	**8 (40%)**	**9 (45%)**	**3 (15%)**	**1.75**
2.1	6	-	6 (100%)	-	2
2.2	5	1 (20%)	4 (80%)	-	1.8
2.3	7	3 (43%)	4 (57%)	-	1.57
2 overall	**18**	**4 (22%)**	**14 (78%)**	**-**	**1.78**
3.1	4	1 (25%)	3 (75%)	-	1.75
3.2	6	1 (17%)	4 (67%)	1 (17%)	2
3.3	7	1 (14%)	6 (86%)	-	1.86
3.4	8	3 (38%)	5 (63%)	-	1.63
3.5	5	-	-	5 (100%)	3
3 overall	**30**	**6 (20%)**	**18 (60%)**	**6 (20%)**	**2**
4.1	6	-	5 (83%)	1 (17%)	2.17
4.2	7	-	7 (100%)	-	2
4.3	5	-	1 (20%)	4 (80%)	2.8
4.4	9	3 (33%)	6 (67%)	-	1.67
4.5	5	-	2 (40%)	3 (60%)	2.6
4 overall	**32**	**3 (9%)**	**21 (66%)**	**8 (25%)**	**2.16**
5.1	7	1 (14%)	6 (86%)	-	1.86
5.2	7	-	5 (71%)	2 (29%)	2.29
5.3	7	4 (57%)	3 (43%)	-	1.43
5.4	6	3 (50%)	2 (33%)	1 (17%)	1.67
5.5	3	-	1 (33%)	2 (67%)	2.67
5 overall	**30**	**8 (27%)**	**17 (57%)**	**5 (17%)**	**1.9**
ALL	**130**	**29 (22%)**	**79 (61%)**	**22 (17%)**	**1.95**

As shown in Table 3, across all standards and the 130 objectives, the distribution of DOK levels was 22% (29) at Level 1, 61% (79) at Level 2, and 17% (22) at Level 3, for an average DOK of 1.95. For each standard, the majority of objectives were assigned DOK Level 2. The 20 objectives in NAEP Standard 1, “Number properties and operations,” were assigned DOK Level 1 (40%), Level 2 (45%), or Level 3 (15%), for an average DOK level of 1.75. The 18 objectives in Standard 2, “Measurement,” were divided between DOK Level 1 (22%) and Level 2 (78%), for an average DOK level of 1.78. The 30 objectives in Standard 3, “Geometry,” were assigned DOK Level 1 (20%), Level 2 (60%), or Level 3 (20%), for an average DOK level of 2. The 32 objectives in Standard 4, “Data analysis, statistics, and probability,” were assigned DOK Level 1 (9%), Level 2 (66%), or Level 3 (25%), for an average DOK level of 2.16. The 30 objectives in Standard 5, “Algebra,” were assigned DOK Level 1 (27%), Level 2 (57%), or Level 3 (17%), for an average DOK level of 1.9. The two standards with the highest percentage of Level 3 objectives and the highest average overall DOK were “Data analysis, statistics, and probability” and “Geometry.”

DOK Levels of the ACCUPLACER Framework

The ACCUPLACER framework used in this study consists of three standards, 12 goals, and 87 objectives. As described earlier, because the objectives in the specifications contain content topics but do not make explicit the intended level of application of knowledge and skills, the ACCUPLACER specifications were not coded for DOK.

DOK Levels of the Test Items

Panelists assigned each item a DOK rating, independent of any content alignment. Because panelists were not required to reach consensus on the DOK values of items, these ratings were not consensus ratings, and interrater agreement for DOK is addressed in Table 2. The average DOK levels of the NAEP items in the short-form set of 42 items used for the NAEP-to-NAEP study were 1.65 for Panel 1 and 1.64 for Panel 2. The average DOK levels of the NAEP items in the complete set of 164 items used for the NAEP-to-ACCUPLACER study were 1.80 for Panel 1 and 1.88 for Panel 2. The average DOK levels for the ACCUPLACER items in the short-form set of 45 items were 1.46 for Panel 1 and 1.50 for Panel 2. The average DOK levels of the ACCUPLACER items in the complete set of 105 items used for the ACCUPLACER-to-NAEP study were 1.43 for Panel 1 and 1.44 for Panel 2. While the average DOK level of the items in the 42-item short-form selected for the NAEP-to-NAEP alignment was slightly lower than the average DOK level of the items in the complete pool of 164 items, both short-form samples appeared to be representative in the sampled areas. The comparison of the DOK levels of the test items with the DOK levels of the NAEP objectives to which they align is addressed in the depth-of-knowledge consistency analyses for Sub-Studies 1 and 2 later in this section. The range of DOK levels of the test items aligned to the ACCUPLACER framework is addressed in the analyses for Sub-Studies 3 and 4.

Alignment Results by Sub-Study

The alignment results of each sub-study are presented in the following sections. As discussed in Section II of this report, the order in which the sub-studies were conducted was modified so that each assessment would be coded to its own framework prior to being coded to the other assessment's framework. For consistency with the design document and to emphasize alignment by framework, the results are presented here in the following order (full WAT reports by sub-study are provided in Appendices J–M; panelist responses to assessment framework debrief questionnaires are provided in Appendices N and O):

- Sub-Study 1. NAEP Items (Short Version) to NAEP Framework
- Sub-Study 2. ACCUPLACER Items to NAEP Framework
- Sub-Study 3. ACCUPLACER Items (Short Version) to ACCUPLACER Framework
- Sub-Study 4. NAEP Items to ACCUPLACER Framework

Sub-Study 1—NAEP Items (Short Version) to NAEP Framework

In Sub-Study 1, panelists evaluated the alignment between the NAEP items and the NAEP framework. A short-form sample of 42 items was analyzed. The results of Sub-Study 1 are presented in Tables 4–8.

Table 4 displays the numbers of items reviewed that were determined to be codable or uncodable. For an item to be codable, at least one reviewer must have coded it to an objective. For an item to be uncodable, all reviewers must have rated it uncodable, that is, not aligned to any objective.

Table 4. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 42

	Panel 1	**Panel 2**
Codable items	42	42
Uncodable items	0	0
Total assessment items	42	42

As shown in Table 4, all 42 items were coded to at least one objective.

Each time a panelist coded an item to an objective was considered one "hit." Mean hits are calculated by dividing the number of hits by the number of panelists. Table 5 displays the numbers and percentages of mean hits by each panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 5. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 42

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	43.63	100%	42.75	100%
Uncodable	0.13	0%	0.00	0%
Total	43.75		42.75	

For the 42 items, the mean hits for the two panels were 43.63 and 42.75. The numbers in the range exceed 42 because some items were coded to multiple objectives by one or more panelists. One panelist in Panel 1 assigned an uncodable rating to one item.

Table 6 shows the categorical concurrence based on the counts of items that were coded to each of the five standards in terms of mean hits, percentage of total hits, and percentage of hits adjusted for items that were determined to be uncodable for each panel. For this sub-study, since no items were identified as uncodable, the percentage of total hits and the adjusted percentage are the same.

Table 6. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 42

	Panel 1			Panel 2		
Standards	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
1 – Number properties and operations	8.50	19%	19%	8.00	19%	19%
2 – Measurement	7.75	18%	18%	6.63	15%	15%
3 – Geometry	6.88	16%	16%	7.13	17%	17%
4 – Data analysis, statistics, and probability	4.50	10%	10%	5.00	12%	12%
5 – Algebra	16.00	37%	37%	16.00	37%	37%
Total	43.63	100%	100%	42.75	100%	100%

Percentages in table may not sum to 100% due to rounding.

All NAEP standards received hits from NAEP items in the short-version subset, with a distribution as indicated in Table 6. Of the five standards, Standard 5, "Algebra," received the greatest number of mean hits in both panels (16.00 in each), making up 37% of the item set for each panel. Standard 4, "Data analysis, statistics, and probability," received the fewest mean hits in both panels: 4.50 mean hits in Panel 1 (10% of total hits) and 5.00 mean hits in Panel 2 (12% of total hits). As mentioned in Section II, it should be noted that, although it was considered sufficiently representative across a number of characteristics and small enough to allow for completion of coding, the short-version item sample was overrepresentative of "Number, properties, and operations" by approximately 9 percentage points (4–5 items) and underrepresentative of "Data analysis, statistics, and probability" by approximately 12 percentage points (5 items).

Reporting categorical concurrence in terms of mean hits and percentage of hits at a finer grain size, Table 7 displays the numbers and percentages of mean hits to objectives. Percentages for this table are reported as the percentage of total hits.

Table 7. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 42

			Panel 1		Panel 2	
Standards	**Goals**	**Objectives**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
1	1.1	1.1.d	0.63	1	1	2
		1.1.f	1	2	1	2
		1.1.g	0	0	0	0
		1.1.i	0.13	0	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
	1.2	1.2.b	0.25	1	0.25	1
		1.2.c	0.13	0	0.88	2
		1.2.d	0	0	0	0
	1.3	1.3.a	0	0	0.13	0
		1.3.b	2.25	5	1.13	3
		1.3.c	1	2	0.75	2
		1.3.d	0	0	0	0
		1.3.f	0.75	2	0.63	1
		1.3	0.13	0	0	0
	1.4	1.4.c	0	0	0	0
		1.4.d	1	2	1	2
		1.4	0	0	0.13	0
	1.5	1.5.c	0.63	1	1	2
		1.5.d	0	0	0	0
		1.5.e	0.13	0	0.13	0
		1.5.f	0	0	0	0
	1.6	1.6.a	0.25	1	0	0
		1.6.b	0.25	1	0	0
2	2.1	2.1.b	0	0	0	0
		2.1.c	0.25	1	0	0
		2.1.d	0.38	1	0	0
		2.1.f	0.63	1	0.38	1
		2.1.h	0	0	0	0
		2.1.i	1.5	3	0.75	2
	2.2	2.2.a	0.88	2	0.88	2
		2.2.b	1.13	3	0.63	1
		2.2.d	0	0	0.13	0
		2.2.e	0.13	0	0.25	1
		2.2.f	2.13	5	2.63	6
	2.3	2.3.a	0.13	0	0	0
		2.3.b	0	0	0.13	0
		2.3.c	0.63	1	0.88	2
		2.3.d	0	0	0	0
		2.3.e	0	0	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		2.3.f	0	0	0	0
		2.3.g	0	0	0	0
3	3.1	3.1.c	0	0	0	0
		3.1.d	0.63	1	0.38	1
		3.1.e	0	0	0	0
		3.1.f	0	0	0	0
	3.2	3.2.a	0	0	0	0
		3.2.b	0	0	0	0
		3.2.c	0.5	1	0.25	1
		3.2.d	0.5	1	0.75	2
		3.2.e	0	0	0	0
		3.2.g	0	0	0	0
	3.3	3.3.b	1.5	3	0.5	1
		3.3.c	0	0	0.13	0
		3.3.d	0.63	1	0.75	2
		3.3.e	0	0	0	0
		3.3.f	0.75	2	0.75	2
		3.3.g	0	0	0	0
		3.3.h	0.63	1	0.63	1
	3.4	3.4.a	1	2	1	2
		3.4.b	0	0	0	0
		3.4.c	0	0	1	2
		3.4.d	0	0	0	0
		3.4.e	0	0	0	0
		3.4.f	0	0	0	0
		3.4.g	0	0	0	0
		3.4.h	0.75	2	1	2
	3.5	3.5.a	0	0	0	0
		3.5.b	0	0	0	0
		3.5.c	0	0	0	0
		3.5.d	0	0	0	0
		3.5.e	0	0	0	0
4	4.1	4.1.a	1	2	1.63	4
		4.1.b	0.25	1	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		4.1.c	0.13	0	0	0
		4.1.d	1.13	3	0	0
		4.1.e	0	0	0	0
		4.1.f	0	0	0	0
	4.2	4.2.a	0.13	0	0.25	1
		4.2.b	0	0	0	0
		4.2.c	0	0	0	0
		4.2.d	0.63	1	0.13	0
		4.2.e	0	0	0	0
		4.2.f	0	0	0	0
		4.2.g	0	0	0	0
	4.3	4.3.a	0	0	0	0
		4.3.b	0	0	0	0
		4.3.c	0.25	1	0.88	2
		4.3.d	0	0	0	0
		4.3.e	0	0	0	0
	4.4	4.4.a	0.13	0	0	0
		4.4.b	0.25	1	0.13	0
		4.4.c	0	0	0.13	0
		4.4.d	0	0	0	0
		4.4.e	0	0	0	0
		4.4.h	0.25	1	0.75	2
		4.4.i	0	0	0	0
		4.4.j	0.38	1	0.13	0
		4.4.k	0	0	0	0
	4.5	4.5.a	0	0	1	2
		4.5.b	0	0	0	0
		4.5.c	0	0	0	0
		4.5.d	0	0	0	0
		4.5.e	0	0	0	0
5	5.1	5.1.a	1.38	3	1.88	4
		5.1.b	1.75	4	1	2
		5.1.e	1.75	4	3.38	8
		5.1.g	0	0	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		5.1.h	0.13	0	0	0
		5.1.i	0.5	1	0	0
		5.1.j	0	0	0	0
	5.2	5.2.a	0.88	2	0.75	2
		5.2.b	0	0	0	0
		5.2.d	0.13	0	0.13	0
		5.2.e	0.75	2	0.63	1
		5.2.f	1.25	3	0.75	2
		5.2.g	0	0	0	0
		5.2.h	0.13	0	0	0
		5.2	0.38	1	0.13	0
	5.3	5.3.b	0.63	1	0	0
		5.3.c	1	2	0.88	2
		5.3.d	0.75	2	1.13	3
		5.3.e	1.38	3	1	2
		5.3.f	0.38	1	1	2
		5.3.g	0	0	0	0
		5.3.h	0.13	0	0	0
	5.4	5.4.a	1	2	1	2
		5.4.c	0	0	0.25	1
		5.4.d	0	0	0.13	0
		5.4.e	0	0	0	0
		5.4.f	0.13	0	0	0
		5.4.g	0	0	0	0
	5.5	5.5.a	1.13	3	1.25	3
		5.5.b	0.5	1	0.75	2
		5.5.c	0	0	0	0

Percentages in table may not sum to 100% due to rounding.

As shown in Table 7, 72 of the 130 objectives received one or more hits; additionally, three goals received at least one hit at the goal level. The two objectives with the greatest number of mean hits were:

- 2.2.f, "Construct or solve problems involving scale drawings" (2.13 mean hits in Panel 1 and 2.63 mean hits in Panel 2)
- 5.1.e, "Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or *trigonometric functions from tables, graphs, or equations[15]" (1.75 mean hits in Panel 1 and 3.38 mean hits in Panel 2)

The majority of panelists across both panels identified three items that were aligned to Objective 2.2.f and two items that were aligned to Objective 5.1.e.

The three items coded to Objective 2.2.f belong to a set of four items clustered around a stimulus and assess skills related to scale drawings. These four items also received codes for other objectives in Goal 2.2, "Systems of measurement," as well as for objectives in Goal 2.1, "Measuring physical attributes," indicating a possible overlap in item content or objective. However, across both panels, the majority of the panelists coded three of these four items to Objective 2.2.f.

The items coded to Objective 5.1.e belong to a set of four individual items that are distributed among the set of 42 items (in other words, that are not clustered around a single stimulus). The coding of these four items by the panelists varied somewhat. Almost all the codes for these four items were within Goals 5.1, "Patterns, relations, and functions," or 5.2, "Algebraic representations"; however, across both panels, the majority of the hits were to Objective 5.1.e.

Within "Geometry," objectives within Goals 3.3 and 3.4 received the majority of hits: objectives within Goal 3.3, "Relationships between geometric figures," received 3.51 and 2.76 mean hits in Panels 1 and 2, respectively, and objectives within Goal 3.4, "Position, direction, and coordinate geometry," received 1.75 and 3.00 mean hits in Panels 1 and 2, respectively. The objectives listed under Goals 3.3 and 3.4 tend to be broad in scope, and items aligned to them can address a range of skills. However, some items were found to align to other goals in this standard, as well. Across the goals for the other two standards—"Number properties and operations" and "Data analysis, statistics, and probability"—the coding of items was more evenly distributed.

As shown in Table 7, of the 130 objectives, 90 in Panel 1 and 93 in Panel 2 received two or fewer hits (0 and .25 mean hits). Two of the standards—"Geometry" and "Data analysis, statistics, and probability"—had the greatest number of objectives receiving two or fewer hits (21 of 30 in each panel for "Geometry," and 28 of 32 in each panel for "Data analysis, statistics, and probability"). However, any conclusions drawn from the data must take into account the large number of objectives (130) relative to the number of items (42) in this study.

[15] "Some of the grade 12 objectives are marked with an "*." This denotes objectives that describe mathematics content beyond that typically taught in a standard 3-year course of study (the equivalent of 1 year of geometry and 2 years of algebra). Therefore, these objectives will be selected less often than the others for inclusion on the assessments" (National Assessment Governing Board, 2008).

The comments recorded by the panelists in debrief questionnaires expressed the observation that some important objectives were not addressed. For example, panelists commented that there were no items coded to Goal 3.5, “Mathematical reasoning in geometry,” and that there were fewer items than expected coded to “Data analysis, statistics, and probability.” These identified gaps are likely attributable to content gaps in the short-version item sample used. When assigning objectives to items, an item can possibly be aligned to more than one objective. For example, one item involved operations on decimal numbers and also assessed estimation and place value. The coding of this item to Objective 1.3.b by four panelists in Panel 1 reflects their judgment that the item is primarily about operations, while the coding of this item to Objective 1.2.c by the six panelists in Panel 2 reflects their judgment that the item is primarily about estimation.

The findings also highlight the overlap in the meaning of some objectives. For example, one item was coded to Objective 5.1.b by five panelists in Panel 1; the same item was coded to Objective 5.1.a by seven panelists in Panel 2. Objective 5.1.a is about “describing . . . geometric progressions,” whereas Objective 5.1.b is about “expressing . . . exponential functions . . . in explicit form given a . . . verbal description.” This item could reasonably be coded to either objective, since it includes both a geometric progression and the exponential rules for that progression.

Table 8 displays the summary of alignment levels on the four content focus criteria for the alignment study: categorical concurrence, depth-of-knowledge consistency, range of knowledge, and balance of representation. The values in the table are intended to be descriptive only. For comparison purposes, asterisks are used to denote values considered “Weak” or “No” according to the typical WAT threshold values.

Table 8. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Eight Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 42

	Alignment Criteria							
	Categorical Concurrence (mean hits)		**Depth-of-Knowledge Consistency** (% of hits at or above level of standard)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
1 – Number properties and operations	8.5	8	86	72	33**	38**	0.86	0.97
2 – Measurement	7.75	6.62	70	74	31**	26**	0.82	0.78
3 – Geometry	6.88	7.12	53	54	20**	23**	0.92	0.99
4 – Data analysis, statistics, and probability	4.5**	5**	59	53	14**	14**	0.98	0.91
5 – Algebra	16	16	75	72	38**	39**	0.83	0.81

One asterisk (*) indicates that the standard would weakly meet the alignment criterion according to the typical WAT threshold values. Two asterisks (**) indicate that the standard would not meet the alignment criterion according to the typical WAT threshold values.

Of the 42 NAEP assessment items analyzed, all (42) were found to match or "hit" objectives. Using the typical WAT threshold value of six mean hits, categorical concurrence was met for Standard 1, "Number properties and operations," with 8.5 and 8 mean hits to the standard. As shown in Table 8, categorical concurrence was also met for Standard 2, "Measurement," with 7.75 and 6.62 mean hits to the standard, and Standard 3, "Geometry," with 6.88 and 7.12 mean hits to the standard. Standard 5, "Algebra," received the most hits, with 16 mean hits to the standard by each panel. Categorical concurrence was not met for Standard 4, "Data analysis, statistics, and probability," with 4.5 and 5 mean hits to the standard. However, it is important to note that there was a distribution of alignment across the standards, and in a larger item pool there would be a greater chance of meeting this numerical threshold.

The depth-of-knowledge consistency criterion was met for all five standards, with more than 50% of the items at or above the DOK level of the standard to which they aligned. The two standards rated lowest in terms of depth-of-knowledge consistency were "Geometry," with 53% and 54% of hits for items at or above the DOK level of the objective, and "Data analysis, statistics, and probability," with 59% and 53% of hits for items at or above the DOK level of the objective. The objectives under these two standards had the highest DOK ratings of the five standards.

Using the typical WAT threshold values, the range of knowledge criterion was not met for any standard. Panelists found that between 14% and 39% of the objectives in each standard were assessed in the item sample. This is largely due to the relatively small number of items being

aligned to a large number of objectives. Given the full item pool, it would be expected that a greater range of objectives would receive hits.

The balance of representation criterion was met for all five standards. That is, item alignments were well distributed among those objectives in each standard that received hits.

Sub-Study 2—ACCUPLACER Items to NAEP Framework

In Sub-Study 2, panelists evaluated the alignment between the ACCUPLACER items and the NAEP framework. One form from each of three tests was combined for a total of 105 items. The results of Sub-Study 2 are presented in Tables 9–13.

Table 9 displays the numbers of items reviewed that were determined to be codable or uncodable. For an item to be codable, at least one reviewer must have coded it to an objective. For an item to be uncodable, all reviewers must have rated it uncodable, that is, not aligned to any objective.

Table 9. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 105

	Panel 1	**Panel 2**
Codable items	105	105
Uncodable items	0	0
Total assessment items	105	105

As shown in Table 9, all 105 ACCUPLACER items were coded to at least one objective.

Each time a panelist coded an item to an objective was considered one "hit." Mean hits are calculated by dividing the number of hits by the number of panelists. Table 10 displays the numbers and percentages of mean hits by each panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 10. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 105

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	107.13	100%	105.00	99%
Uncodable	0.13	0%	1.13	1%
Total	107.25		106.13	

For the 105 items, the mean hits for the two panels were 107.25 and 106.13. The numbers in the range exceed 105 because some items were coded to multiple objectives by one or more panelists. One panelist in Panel 1 assigned an uncodable rating to one item. One or more panelists in Panel 2 assigned an uncodable rating to three items.

Table 11 shows the categorical concurrence based on the counts of items that were coded to each of the five standards in terms of mean hits, percentage of total hits, and percentage of hits adjusted for items that were determined to be uncodable for each panel. For this sub-study, since

no items were identified as uncodable, the percentage of total hits and the adjusted percentage are the same.

Table 11. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 105

	Panel 1			**Panel 2**		
Standards	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
1 – Number properties and operations	40.38	38%	38%	40.63	39%	38%
2 – Measurement	3.25	3%	3%	4.63	4%	4%
3 – Geometry	1.00	1%	1%	0.38	0%	0%
4 – Data analysis, statistics, and probability	2.50	2%	2%	2.38	2%	2%
5 – Algebra	60.00	56%	56%	57.00	54%	54%
Total	107.13	100%	100%	105.00	100%	99%

Percentages in table may not sum to 100% due to rounding.

All NAEP standards received hits from the ACCUPLACER items, although the large majority of items were aligned to Standard 1, “Number properties and operations,” and Standard 5, “Algebra.” Of the five standards, “Algebra” received the majority of mean hits in both panels: 60.00 mean hits in Panel 1 (56% of total hits) and 57.00 mean hits in Panel 2 (54% of total hits). “Number properties and operations” also received a substantial number of mean hits in both panels: 40.38 mean hits in Panel 1 (38% of total hits) and 40.63 mean hits in Panel 2 (39% of total hits). Combined, Standard 2, “Measurement”; Standard 3, “Geometry”; and Standard 4, “Data analysis, statistics, and probability” received less than 10% of the total hits. “Geometry” received the fewest mean hits in both panels: 1.00 mean hits in Panel 1 (1% of total hits) and 0.38 mean hits in Panel 2 (0% of total hits).

In comparison with the baseline alignment distribution to the NAEP framework in Sub-Study 1, the ACCUPLACER items had approximately 20 percentage points greater emphasis on “Number properties and operations” and “Algebra” than did the short-version subset of the NAEP items. The ACCUPLACER items had 11–15 percentage points less emphasis on “Measurement,” 15–17 percentage points less emphasis on “Geometry,” and 8–10 percentage points less emphasis on “Data analysis, statistics, and probability” than did the short-version subset of the NAEP items.

Reporting categorical concurrence in terms of mean hits and percentage of hits at a finer grain size, Table 12 displays the numbers and percentages of mean hits to objectives. Percentages for this table are reported as the percentage of total hits.

Table 12. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 105

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
1	1.1	1.1.d	1.63	2	1.88	2
		1.1.f	0	0	0	0
		1.1.g	0.63	1	0.75	1
		1.1.i	4.25	4	4.63	4
	1.2	1.2.b	0.38	0	0.13	0
		1.2.c	0	0	0.25	0
		1.2.d	0.88	1	0.75	1
		1.2	0.38	0	0	0
	1.3	1.3.a	1	1	0.63	1
		1.3.b	19	18	20.25	19
		1.3.c	0.63	1	0.25	0
		1.3.d	0	0	0.13	0
		1.3.f	5.75	5	5.25	5
	1.4	1.4.c	1.88	2	1.63	2
		1.4.d	2.38	2	3.5	3
		1.4	0.88	1	0.13	0
	1.5	1.5.c	0.13	0	0.25	0
		1.5.d	0.13	0	0	0
		1.5.e	0	0	0	0
		1.5.f	0.5	0	0.25	0
	1.6	1.6.a	0	0	0	0
		1.6.b	0	0	0	0
2	2.1	2.1.b	0	0	0	0
		2.1.c	0	0	0	0
		2.1.d	0	0	0	0
		2.1.f	0.13	0	1.13	1
		2.1.h	0	0	0	0
		2.1.i	0.13	0	0.25	0
		2.1	0.25	0	0	0
	2.2	2.2.a	0	0	0	0
		2.2.b	0	0	1	1

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		2.2.d	0	0	0	0
		2.2.e	0	0	0	0
		2.2.f	0	0	0	0
	2.3	2.3.a	0.13	0	0	0
		2.3.b	0.38	0	0.25	0
		2.3.c	1.63	2	1.75	2
		2.3.d	0	0	0	0
		2.3.e	0.13	0	0	0
		2.3.f	0.13	0	0.13	0
		2.3.g	0	0	0	0
		2.3	0.38	0	0.13	0
3	3.1	3.1.c	0	0	0	0
		3.1.d	0	0	0	0
		3.1.e	0	0	0	0
		3.1.f	0	0	0	0
	3.2	3.2.a	0	0	0	0
		3.2.b	0	0	0	0
		3.2.c	0	0	0	0
		3.2.d	0	0	0	0
		3.2.e	0	0	0	0
		3.2.g	0	0	0	0
	3.3	3.3.b	0.13	0	0.13	0
		3.3.c	0	0	0	0
		3.3.d	0	0	0	0
		3.3.e	0	0	0.13	0
		3.3.f	0	0	0.13	0
		3.3.g	0.13	0	0	0
		3.3.h	0	0	0	0
	3.4	3.4.a	0.75	1	0	0
		3.4.b	0	0	0	0
		3.4.c	0	0	0	0
		3.4.d	0	0	0	0
		3.4.e	0	0	0	0
		3.4.f	0	0	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		3.4.g	0	0	0	0
		3.4.h	0	0	0	0
	3.5	3.5.a	0	0	0	0
		3.5.b	0	0	0	0
		3.5.c	0	0	0	0
		3.5.d	0	0	0	0
		3.5.e	0	0	0	0
4	4.1	4.1.a	0.88	1	0.88	1
		4.1.b	0	0	0	0
		4.1.c	0	0	0	0
		4.1.d	0	0	0	0
		4.1.e	0	0	0	0
		4.1.f	0	0	0	0
	4.2	4.2.a	0.75	1	0.63	1
		4.2.b	0	0	0	0
		4.2.c	0	0	0	0
		4.2.d	0	0	0	0
		4.2.e	0.13	0	0	0
		4.2.f	0	0	0	0
		4.2.g	0	0	0	0
	4.3	4.3.a	0	0	0	0
		4.3.b	0	0	0	0
		4.3.c	0	0	0	0
		4.3.d	0	0	0	0
		4.3.e	0	0	0	0
	4.4	4.4.a	0	0	0	0
		4.4.b	0	0	0	0
		4.4.c	0	0	0.13	0
		4.4.d	0	0	0	0
		4.4.e	0.75	1	0.75	1
		4.4.h	0	0	0	0
		4.4.i	0	0	0	0
		4.4.j	0	0	0	0
		4.4.k	0	0	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
	4.5	4.5.a	0	0	0	0
		4.5.b	0	0	0	0
		4.5.c	0	0	0	0
		4.5.d	0	0	0	0
		4.5.e	0	0	0	0
5	5.1	5.1.a	0.25	0	0.75	1
		5.1.b	0.13	0	0	0
		5.1.e	5.13	5	3.5	3
		5.1.g	0	0	0	0
		5.1.h	0.75	1	1.13	1
		5.1.i	1.13	1	1.13	1
		5.1.j	0.88	1	0.75	1
		5.1	0.38	0	0	0
	5.2	5.2.a	4.75	4	2.88	3
		5.2.b	0.25	0	0	0
		5.2.d	1	1	2	2
		5.2.e	0	0	0	0
		5.2.f	0	0	0	0
		5.2.g	0	0	0	0
		5.2.h	0.5	0	0.88	1
	5.3	5.3.b	2.13	2	2.63	3
		5.3.c	16.63	15	20	19
		5.3.d	4	4	1.13	1
		5.3.e	3.25	3	2.63	3
		5.3.f	1.5	1	1	1
		5.3.g	0	0	0.13	0
		5.3.h	0.25	0	0.38	0
		5.3	0.13	0	0	0
	5.4	5.4.a	10.13	9	9.13	9
		5.4.c	1.88	2	2.5	2
		5.4.d	4.13	4	3.63	3
		5.4.e	0	0	0.13	0
		5.4.f	0.75	1	0.75	1
		5.4.g	0	0	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		5.4	0.13	0	0	0
	5.5	5.5.a	0	0	0	0
		5.5.b	0	0	0	0
		5.5.c	0	0	0	0

Percentages in table may not sum to 100% due to rounding.

As shown in Table 12, 56 of the 130 objectives received one or more hits in at least one panel; additionally, items were coded directly to seven goals. The two objectives with the greatest number of mean hits received over two-thirds of the total hits:

- 1.3.b, "Perform arithmetic operations with real numbers, including common irrational numbers" (19 mean hits in Panel 1 and 20.25 mean hits in Panel 2)
- 5.3.c, "Perform basic operations, using appropriate tools, on algebraic expressions including polynomial and rational expressions" (16.63 mean hits in Panel 1 and 20 mean hits in Panel 2)

The majority of panelists across both panels identified 19 items that were aligned to Objective 1.3.b and 18 items that were aligned to Objective 5.3.c.

The items coded to Objective 1.3.b can be characterized as assessing computation, with an emphasis on assessing the mastery of algorithmic skills. The items identified as coded to Objective 5.3.c are also algorithmic in nature, but involve one or more variables. Together, the two objectives received a large percentage of the total hits across both panels (33% in Panel 1 and 38% in Panel 2). Of the 37 items found by panelists to align to the two objectives, none involved a real-world context.

As shown in Table 12, the following 12 objectives had over two mean hits in at least one panel:

- 1.1.i, "Order or compare real numbers, including very large and very small real numbers"
- 1.3.b, "Perform arithmetic operations with real numbers, including common irrational numbers"
- 1.3.f, "Solve application problems involving numbers, including rational and common irrationals"
- 1.4.d, "Solve multistep problems involving percentages, including compound percentages"
- 5.1.e, "Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or *trigonometric functions from tables, graphs, or equations"
- 5.2.a, "Create and translate between different representations of algebraic expressions, equations, and inequalities (e.g., linear, quadratic, exponential, or *trigonometric) using symbols, graphs, tables, diagrams, or written descriptions"
- 5.3.b, "Write algebraic expressions, equations, or inequalities to represent a situation"
- 5.3.c, "Perform basic operations, using appropriate tools, on algebraic expressions including polynomial and rational expressions"

- 5.3.d, “Write equivalent forms of algebraic expressions, equations, or inequalities to represent and explain mathematical relationships”
- 5.3.e, “Evaluate algebraic expressions including polynomials and rational expressions”
- 5.4.a, “Solve linear, rational, or quadratic equations or inequalities, including those involving absolute value”
- 5.4.d, “Solve (symbolically or graphically) a system of equations or inequalities and recognize the relationship between the analytical solution and graphical solution”

As can be seen from the preceding list, three of the six goals in “Number properties and operations,” and four of the five goals in “Algebra,” received a majority of the total hits (over 90% of the total hits). These goals can be characterized as basic skills necessary to compute and solve problems involving real numbers and perform algebraic manipulations. Important objectives from “Number properties and operations” and “Algebra” not included in the list involve certain higher-order thinking and reasoning skills (i.e. objectives under Goal 1.2, “Estimation”; Goal 1.5, “Properties of numbers and operations”; Goal 1.6, “Mathematical reasoning using number”; and Goal 5.5, “Mathematical reasoning in algebra”). Of the 130 objectives, a total of 82 objectives in Panel 1 and 87 objectives in Panel 2 received no hits.

Of the five standards, three received very few hits. Each of these standards (“Measurement,” “Geometry,” and “Data analysis, statistics, and probability”), combined, received less than 10% of the total hits. Of the 83 objectives in these three standards, only one objective received a significant number of mean hits (2.3.c, “Use the definitions of sine, cosine, and tangent as ratios of sides in a right triangle to solve problems about length of sides and measure of angles”). Panelists identified two items in the set of 105 items that align to Objective 2.3.c.

The comments recorded by the panelists expressed the observation that some important objectives were not addressed. For example, panelists commented that there were very few items coded to “Measurement,” “Geometry,” and “Data analysis, statistics, and probability.” In addition, panelists noticed that there were no items at DOK Level 3 and that most items were at Level 1. Most of these items appear to address the skills of computation and algebraic manipulation.

When assigning objectives to items, an item can possibly be aligned to more than one objective. For example, an item involving operations on decimal numbers and also assessing ordering and comparing decimal numbers was coded by most panelists across both panels to Objective 1.3.b, “Perform arithmetic operations with real numbers, including common irrational numbers.” A few panelists, however, coded the item to Objective 1.1.i, “Order or compare real numbers, including very large and very small real numbers.” At least two panelists in each panel recorded a secondary code for this item.

Table 13 displays the summary of alignment levels on the four content focus criteria for the alignment study: categorical concurrence, depth-of-knowledge consistency, range of knowledge, and balance of representation. The values in the table are intended to be descriptive only. For comparison purposes, asterisks are used to denote values considered “Weak” or “No” according to the typical WAT threshold values.

Table 13. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Eight Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 105

	Alignment Criteria							
	Categorical Concurrence (mean hits)		**Depth-of-Knowledge Consistency** (% of hits at or above level of standard)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
1 – Number properties and operations	40.38	40.62	89	89	52	46*	0.54**	0.58**
2 – Measurement	3.25**	4.62**	74	81	13**	19**	0.9	0.88
3 – Geometry	1**	0.38**	76	100	3**	1**	0.73	0.25**
4 – Data analysis, statistics, and probability	2.5**	2.38**	75	85	8**	7**	1	0.88
5 – Algebra	60	57	80	86	52	55	0.61*	0.58**

One asterisk (*) indicates that the standard would weakly meet the alignment criterion according to the typical WAT threshold values. Two asterisks (**) indicate that the standard would not meet the alignment criterion according to the typical WAT threshold values.

Of the 105 ACCUPLACER assessment items analyzed, all (105) were found to match or "hit" objectives. Using the typical WAT threshold value of six mean hits, categorical concurrence was met for Standard 1, "Number properties and operations," with 40.38 (Panel 1) and 40.62 (Panel 2) mean hits to the standard. Categorical concurrence was also met for Standard 5, "Algebra," with 57 (Panel 2) and 60 (Panel 1) mean hits to the standard. Categorical concurrence was not met for "Measurement," "Geometry," or "Data analysis, statistics, and probability," with fewer than 4.62 mean hits to each standard. In comparing these results to the baseline alignment in Sub-Study 1, it is important to consider the difference in the number of items reviewed and compare the relative proportional distribution between the two tests. The proportionally very low numbers of hits for "Measurement," "Geometry," and "Data analysis, statistics, and probability" should be taken into consideration in interpreting depth-of-knowledge consistency, range of knowledge, and balance of representation.

The depth-of-knowledge consistency criterion was met for all five standards, with more than 50% of the items at or above the DOK level of the standard to which they aligned. Most of the items in the set of 105 items were rated at DOK Level 1, and all were rated at DOK Level 1 or 2. In addition, the objectives with the greatest numbers of items aligned to them were rated at DOK Level 1 or 2. Discrepancies of greater than five percentage points existed between the panels for this criterion for "Measurement," "Geometry," and "Data analysis, statistics, and probability." This was identified during cross-panel adjudication. It was determined that differences were not systematic, particularly given the very small numbers of mean hits for these standards. The facilitators identified the items that had the greatest differences in DOK coding by the two

panels. The differences discussed within and across panels tended to focus on the interpretation of grade-level appropriateness and the extent to which a student had to apply reasoning skills in items (with or without a context) involving multiple steps and/or multiple concepts. As students move up through the grades, items that might have been more cognitively complex for students in lower grades can become more routine, so that even, for example, items in context involving two steps or two concepts can be Level 1 if the items require the routine application of simple procedures. Following cross-panel adjudication, panelists tended to become more consistent within and across panels in their coding of DOK.

Range of knowledge was not met for any of the five standards of the NAEP framework. The limited range of knowledge found in the ACCUPLACER items is consistent with differences in the frameworks of the two tests: the ACCUPLACER framework includes many objectives involving number computation and algebraic manipulation, while the NAEP framework includes additional content areas, as well as objectives that involve higher-level thinking and application of concepts. Therefore, one would expect ACCUPLACER to cover a narrower range of concepts than the NAEP objectives require.

The balance of representation criterion is most relevant in this sub-study for “Number properties and operations” and “Algebra,” where the overwhelming majority of hits were. Although a range of objectives received hits, the typical WAT threshold of 0.7 was not met for these two standards, with between 0.54 and 0.61 mean hits per standard. This indicates that the items were clustered around one or two objectives in each standard. In this case, the items were clustered around Objectives 1.3.b, 5.3.c, and 5.4.a. As stated previously, these items tended to be skills-based and algorithmic in nature, qualities that characterize the ACCUPLACER items as a whole.

Sub-Study 3—ACCUPLACER Items to ACCUPLACER Framework

In Sub-Study 3, panelists evaluated the alignment between the ACCUPLACER items and the ACCUPLACER framework. A short-form sample of 45 items was analyzed. The results of Sub-Study 3 are presented in Tables 14–19.

Table 14 displays the number of items reviewed that were determined to be codable or uncodable. For an item to be codable, at least one reviewer must have coded it to an objective. For an item to be uncodable, all reviewers must have rated it uncodable, that is, not aligned to any objective.

Table 14. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework
Assessment items = 45

	Panel 1	**Panel 2**
Codable items	45	45
Uncodable items	0	0
Total assessment items	45	45

As shown in Table 14, all 45 ACCUPLACER items were coded to at least one objective.

Each time a panelist coded an item to an objective was considered one "hit." Mean hits are calculated by dividing the number of hits by the number of panelists. Table 15 displays the numbers and percentages of mean hits by each panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 15. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework
Assessment items = 45

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	46.00	100%	46.88	100%
Uncodable	0.00	0%	0.00	0%
Total	46.00		46.88	

For the 45 ACCUPLACER items, the mean hits for the two panels were 46.00 and 46.88. The numbers in the range exceed 45 because some items were coded to multiple objectives by one or more panelists. No uncodable ratings were assigned.

Table 16 shows the categorical concurrence based on the counts of items that were coded to each of the three standards in terms of mean hits, percentage of total hits, and percentage of hits adjusted for items that were determined to be uncodable for each panel. For this sub-study, since no items were identified as uncodable, the percentage of total hits and the adjusted percentage are the same.

Table 16. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework
Assessment items = 45

Standards	Panel 1			Panel 2		
	Mean Hits	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
A – Arithmetic	15.38	33%	33%	16.50	35%	35%
B – Elementary algebra	18.13	39%	39%	16.88	36%	36%
C – College level math	12.50	27%	27%	13.50	29%	29%
Total	46.00	100%	100%	46.88	100%	100%

Percentages in table may not sum to 100% due to rounding.

All ACCUPLACER standards received hits from ACCUPLACER items in the short-version subset, with a distribution similar to that in the framework. Of the three standards, ACCUPLACER B, "Elementary algebra," received the greatest number of mean hits in both panels: 18.13 mean hits in Panel 1 (39% of total hits) and 16.88 mean hits in Panel 2 (36% of total hits). ACCUPLACER A, "Arithmetic," received the second greatest number of mean hits in both panels: 15.38 mean hits in Panel 1 (33% of total hits) and 16.50 mean hits in Panel 2 (35% of total hits). ACCUPLACER C, "College level math," received the fewest mean hits in both panels: 12.50 mean hits in Panel 1 (27% of total hits) and 13.50 mean hits in Panel 2 (29% of total hits). In general, panelists found that emphasis across the 45 items was fairly evenly distributed across the ACCUPLACER standards.

Reporting categorical concurrence in terms of mean hits and percentage of hits at a finer grain size, Table 17 displays the numbers and percentages of mean hits to objectives. Percentages for this table are reported as the percentage of total hits.

Table 17. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework
Assessment items = 45

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
A	A.1	A.1.a	0.63	1	0.25	1
		A.1.b	1.13	2	1	2
		A.1.c	0.88	2	1	2
		A.1.d	0.25	1	0	0
		A.1.e	0	0	0	0
		A.1.f	0	0	0	0
		A.1.g	0.38	1	0.63	1
		A.1.h	0.88	2	0.75	2
		A.1.i	0.38	1	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
	A.2	A.2.a	0.13	0	0	0
		A.2.b	0.25	1	0.25	1
		A.2.c	0.75	2	0.88	2
		A.2.d	1.25	3	1.13	2
		A.2.e	0.13	0	0	0
		A.2.f	0.75	2	0.88	2
		A.2.g	0.13	0	0.13	0
		A.2.h	1	2	2	4
		A.2.i	0.13	0	0	0
		A.2.j	0.75	2	1	2
		A.2.k	0	0	0	0
		A.2.l	0	0	0	0
	A.3	A.3.a	1.13	2	1	2
		A.3.b	1.13	2	1	2
		A.3.c	1.25	3	1.88	4
		A.3.d	2	4	2.13	5
		A.3.e	0	0	0	0
		A.3.f	0	0	0	0
		A.3.g	0.13	0	0.63	1
B	B.1	B.1.a	0.5	1	0.38	1
		B.1.b	0.75	2	1	2
		B.1.c	1.38	3	1	2
	B.2	B.2.a	2	4	2	4
		B.2.b	0	0	0	0
		B.2.c	2	4	2	4
		B.2.d	1.25	3	1	2
		B.2.e	0	0	0	0
		B.2.f	3.5	8	3.5	7
		B.2.g	0	0	0	0
		B.2.h	0	0	0	0
		B.2.i	1.38	3	1.13	2
		B.2.j	1.13	2	0.63	1
	B.3	B.3.a	2	4	2.13	5
		B.3.b	0.5	1	0.88	2

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		B.3.c	0.13	0	0	0
		B.3.d	0	0	0	0
		B.3.e	0.63	1	1.25	3
		B.3.f	1	2	0	0
C	C.1	C.1.a	0	0	0	0
		C.1.b	0	0	0	0
		C.1.c	0	0	0	0
		C.1.d	0.63	1	1.25	3
		C.1.e	0	0	0	0
		C.1.f	0.63	1	0.25	1
		C.1.g	0.13	0	0.5	1
	C.2	C.2.a	0.38	1	0.88	2
		C.2.b	1	2	1.75	4
		C.2.c	0.38	1	0.38	1
		C.2.d	0	0	0	0
		C.2.e	0	0	0	0
	C.3	C.3.a	0.25	1	0.88	2
		C.3.b	1	2	0.63	1
		C.3.c	0	0	0	0
		C.3.d	0	0	0	0
		C.3.e	0.63	1	1.13	2
	C.4	C.4.a	0	0	0	0
		C.4.b	0	0	0	0
		C.4.c	0	0	0	0
		C.4.d	0	0	0	0
		C.4.e	0.88	2	0.5	1
		C.4.f	0	0	0	0
		C.4.g	0	0	0	0
	C.5	C.5	0.13	0	0	0
		C.5.a	0	0	0	0
		C.5.b	0	0	0	0
		C.5.c	0.38	1	0.13	0
		C.5.d	0.63	1	0.13	0
		C.5.e	0.13	0	0.13	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		C.5.f	0.88	2	0.88	2
		C.5.g	0.75	2	0.75	2
		C.5.h	0.88	2	0.63	1
	C.6	C.6.a	1.13	2	1.13	2
		C.6.b	0.75	2	0.88	2
		C.6.c	0	0	0	0
		C.6.d	0.75	2	0.75	2
		C.6.e	0.13	0	0	0
		C.6.f	0	0	0	0
		C.6.g	0	0	0	0
		C.6.h	0.13	0	0	0

Percentages in table may not sum to 100% due to rounding.

As shown in Table 17, 57 of the 87 objectives received one or more hits in at least one panel; additionally, one item was coded directly to a goal. Two objectives, both within ACCUPLACER B, “Elementary algebra,” had the greatest number of mean hits in both panels:

- B.2.f, “Factoring difference of squares” (3.5 mean hits in both panels)
- B.3.a, “Solving linear equations and inequalities” (2 mean hits in Panel 1 and 2.13 mean hits in Panel 2)

The majority of panelists across both panels identified five items that were aligned to Objective B.2.f and two items that were aligned to Objective B.3.a.

The items coded to Objective B.2.f can be characterized as involving a certain algorithmic procedure that is basic to beginning algebra. The items identified as coded to Objective B.3.a are procedural in nature as well, although they potentially may be assessed in a real-world context.

As stated previously, each of the three standards of the ACCUPLACER framework received approximately equal emphasis across the 45 items in the short-form subset of items. The number of total hits within each standard tended to be distributed evenly as well. Of the 28 objectives in ACCUPLACER A, “Arithmetic,” 22 objectives received one or more hits in at least one panel. Of the 19 objectives in ACCUPLACER B, “Elementary algebra,” 14 objectives received one or more hits. Of the 40 objectives in ACCUPLACER C, “College level math,” 21 objectives received one or more hits.

As shown in Table 17, 30 of the 87 objectives received no hits. Of those 30 objectives receiving no hits, 19 objectives were in “College level math.” Within “College level math,” fewer than one-half of the objectives in two goals (C.1, “Algebraic operations,” and C.4, “Applications and other algebra topics”) received hits. However, any conclusions drawn from the data must take

into account the large number of objectives (87) relative to the number of items (45) in this sub-study.

The comments recorded by the panelists did express the observation that the objectives in Goals C.1 and C.4 were not adequately addressed but acknowledged that the number of items compared to the total number of objectives likely was the reason.

When assigning objectives to items, an item can possibly be aligned to more than one objective. One example of this is an item that involves both operations on decimal numbers and estimation and place value. The coding of the item to Objective A.2.h by 50% of all panelists reflects their judgment that the item is primarily about operations. The remaining panelists chose other objectives, reflecting their judgment that the item is primarily about estimation, which is not mentioned in Objective A.2.h.

Table 18 displays the summary of alignment levels on three of the four content focus criteria for the alignment study: categorical concurrence, range of knowledge, and balance of representation. The ACCUPLACER framework was not able to be coded for depth-of-knowledge consistency. The values in the table are intended to be descriptive only. For comparison purposes, asterisks are used to denote values considered "Weak" or "No" according to the typical WAT threshold values.

Table 18. Summary of Attainment of Acceptable Alignment Level on Three Content Focus Criteria as Rated by Eight Reviewers per Panel—ACCUPLACER Items (Short Version) to ACCUPLACER Framework
Assessment items = 45

	Alignment Criteria					
	Categorical Concurrence (mean hits)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
A – Arithmetic	15.38	16.5	47*	47*	0.88	0.85
B – Elementary algebra	18.12	16.88	59	55	0.8	0.8
C – College level math	12.5	13.5	30**	29**	0.97	0.9

One asterisk (*) indicates that the standard would weakly meet the alignment criterion according to the typical WAT threshold values. Two asterisks (**) indicate that the standard would not meet the alignment criterion according to the typical WAT threshold values.

Of the 45 ACCUPLACER assessment items analyzed, all (45) were found to match objectives, or have "hits." Using the typical WAT threshold value of six mean hits, categorical concurrence was met for ACCUPLACER A, "Arithmetic," with 15.38 (Panel 1) and 16.5 (Panel 2) mean hits to the standard. As shown in Table 18, categorical concurrence was also met for ACCUPLACER B, "Elementary algebra," with 16.88 (Panel 2) and 18.12 (Panel 1) mean hits to the standard, and ACCUPLACER C, "College level math," with 12.5 (Panel 1) and 13.5 (Panel 2) mean hits to the standard.

The range of knowledge criterion was only met above the 50% criterion for "Elementary algebra." Range of knowledge was met weakly for "Arithmetic," and not met for "College level

math." Given the full item pool, it would be expected that a greater range of objectives would receive hits.

The balance of representation criterion was met for all three standards. There are no indications from panelists' comments that any particular objective(s) received more hits than might have been expected. Of the objectives receiving hits, the distribution of items across those objectives was well distributed.

As described earlier, the ACCUPLACER framework could not be coded for DOK. Table 19 shows the range of depth of knowledge of the items aligned to each ACCUPLACER standard.

Table 19. Range of Depth of Knowledge of ACCUPLACER Items Aligned to the ACCUPLACER Framework
Assessment items = 45

	Panel 1							**Panel 2**						
		DOK Level 1		**DOK Level 2**		**DOK Level 3**			**DOK Level 1**		**DOK Level 2**		**DOK Level 3**	
Standards	**Mean Hits**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**
A	15.38	7.63	50	7.75	50	0.00	0	16.50	8.63	52	7.88	48	0.00	0
B	18.12	12.88	71	5.25	29	0.00	0	16.88	12.38	74	4.38	26	0.13	1
C	12.50	4.13	33	8.38	67	0.00	0	13.50	2.25	17	11.25	83	0.00	0
Total	46.00	24.63	51	21.38	49	0.00	0	46.88	23.25	47	23.50	53	0.13	0

As shown in Table 19, almost all items were found to be at DOK Level 1 or 2. The items aligned to ACCUPLACER A, "Arithmetic," were almost evenly divided (50% at each level in Panel 1, and 52% at Level 1 and 48% at Level 2 in Panel 2). Items aligned to ACCUPLACER B, "Elementary algebra," were mostly at Level 1 (71% in Panel 1 and 74% in Panel 2), with the remaining items at Level 2, except for one mean hit at Level 3 assigned by a panelist in Panel 2. Items aligned to ACCUPLACER C, "College level math," were mostly at DOK Level 2 (67% in Panel 1 and 83% in Panel 2), with the remaining items at Level 1. This difference of greater than 5% was identified during cross-panel adjudication. Given the small number of items in the set and the 2–3 mean hits causing the percentage difference, the discrepancy was determined not to be systematic.

Sub-Study 4—NAEP Items to ACCUPLACER Framework

In Sub-Study 4, panelists evaluated the alignment between the NAEP items and the ACCUPLACER framework. All 164 NAEP items were analyzed. The results of Sub-Study 4 are presented in Tables 20–25.

Table 20 displays the number of items reviewed that were determined to be codable or uncodable. For an item to be codable, at least one reviewer must have coded it to an objective. For an item to be uncodable, all reviewers must have rated it uncodable, that is, not aligned to any objective.

Table 20. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 164

	Panel 1	**Panel 2**
Codable items	161	150
Uncodable items	3	14
Total assessment items	164	164

As shown in the table, both panels found a number of NAEP items (3 and 14) to be uncodable to the ACCUPLACER objectives. The specific skills included in the uncodable items are discussed following Table 21.

Each time a panelist coded an item to an objective was considered one "hit." Mean hits are calculated by dividing the number of hits by the number of panelists. Table 21 displays the numbers and percentages of mean hits assigned to items by panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 21. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 164

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	141.50	85%	125.88	76%
Uncodable	24.50	15%	40.50	24%
Total	166.00		166.38	

For the 164 items, the mean hits for the two panels were 141.50 and 125.88. As shown in Table 21, there were a number of uncodable ratings in each panel, comprising 15% and 24% of mean hits in Panel 1 and Panel 2, respectively. While some of these were items found to be uncodable by one or two panelists, 20 items were found to be uncodable by the majority (5 or more of 8) panelists in each panel. Reviewers' comments for specific items identify three main types of items within the items designated as uncodable: NAEP items written to NAEP Standards 2,

“Measurement”; 3, “Geometry”; or 4, “Data analysis, statistics, and probability.” These items, in general, have no comparable sets of objectives within the ACCUPLACER set of objectives.

Table 22 shows the categorical concurrence based on the counts of items that were coded to each of the three standards in terms of mean hits, percentage of total hits, and percentage of hits adjusted for items that were determined to be uncodable for each panel.

Table 22. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 164

	Panel 1			**Panel 2**		
Standards	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
A – Arithmetic	50.00	35%	30%	42.25	34%	25%
B – Elementary algebra	43.88	31%	26%	35.13	28%	21%
C – College level math	47.63	34%	29%	48.50	39%	29%
Total	141.50	100%	85%	125.88	100%	76%

Percentages in table may not sum to 100% due to rounding.

All ACCUPLACER standards received hits from NAEP items, with a distribution similar to that in the framework. Of the three standards, ACCUPLACER A, “Arithmetic,” and ACCUPLACER C, “College level math,” received the greatest number of mean hits in both panels. ACCUPLACER B, “Elementary algebra,” received the fewest mean hits in both panels.

In comparison with the baseline alignment distribution to the ACCUPLACER framework in Sub-Study 3, considering codable items only, the NAEP item pool and the ACCUPLACER short-version subset had very similar degrees of emphasis on “Arithmetic.” The NAEP item pool had approximately 5–11 percentage points less emphasis on “Elementary algebra” and approximately 5–12 percentage points greater emphasis on “College level math” than did the short-version subset of ACCUPLACER items. Taking into account the percentages of total hits adjusted for uncodable items in this sub-study, the distributions across standards between the two sub-studies remain proportionally very similar, but it is important to note that codable hits make up only 85% and 76% of hits in Panels 1 and 2.

Table 23. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 164

			Panel 1		**Panel 2**	
Standards	**Goals**	**Objectives**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
A		A	0.13	0	0	0
	A.1	A.1.a	2.25	2	0.5	0
		A.1.b	0.13	0	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		A.1.c	0.5	0	0.13	0
		A.1.d	0	0	0	0
		A.1.e	0	0	0	0
		A.1.f	0.13	0	0.25	0
		A.1.g	0.63	0	0.25	0
		A.1.h	0	0	0	0
		A.1.i	0.63	0	0.25	0
	A.2	A.2	0.13	0	0	0
		A.2.a	0	0	0	0
		A.2.b	1.25	1	1	1
		A.2.c	0.5	0	0.75	1
		A.2.d	0	0	0	0
		A.2.e	0.38	0	0.75	1
		A.2.f	0	0	0	0
		A.2.g	0.25	0	0.13	0
		A.2.h	0.5	0	0.88	1
		A.2.i	0.38	0	0	0
		A.2.j	0	0	0.13	0
		A.2.k	0	0	0	0
		A.2.l	0	0	0	0
	A.3	A.3.a	8.63	6	9	7
		A.3.b	1.5	1	2	2
		A.3.c	0.38	0	1.5	1
		A.3.d	8.63	6	9.38	7
		A.3.e	2	1	2.25	2
		A.3.f	15.25	11	10.75	9
		A.3.g	5.88	4	2.38	2
B	B.1	B.1.a	0.75	1	0.88	1
		B.1.b	0.25	0	0.38	0
		B.1.c	1.38	1	0.63	0
	B.2	B.2.a	6.63	5	5.5	4
		B.2.b	0.63	0	0.38	0
		B.2.c	1	1	1.13	1
		B.2.d	4	3	3.5	3

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		B.2.e	0.25	0	0.63	0
		B.2.f	0.13	0	0	0
		B.2.g	0.5	0	0.63	0
		B.2.h	0	0	0	0
		B.2.i	1.13	1	0.75	1
		B.2.j	0.13	0	0	0
		B.2	0.38	0	0.88	1
	B.3	B.3.a	2.5	2	3.5	3
		B.3.b	1.13	1	1.75	1
		B.3.c	0.13	0	0.13	0
		B.3.d	5	4	4.5	4
		B.3.e	9.25	7	4.38	3
		B.3.f	8	6	5.63	4
		B.3	0.75	1	0	0
C		C	0.25	0	0	0
	C.1	C.1.a	0.5	0	0.13	0
		C.1.b	0	0	0.13	0
		C.1.c	0	0	0.13	0
		C.1.d	0.25	0	0	0
		C.1.e	0.38	0	1.13	1
		C.1.f	1.5	1	1.13	1
		C.1.g	0	0	0.25	0
		C.1	0.25	0	0	0
	C.2	C.2.a	0.13	0	0.25	0
		C.2.b	1.25	1	1.13	1
		C.2.c	0.38	0	0	0
		C.2.d	2	1	2	2
		C.2.e	0.5	0	0	0
		C.2	0.38	0	0	0
	C.3	C.3.a	3.63	3	3.13	2
		C.3.b	1.75	1	3.38	3
		C.3.c	2.38	2	2.13	2
		C.3.d	0.13	0	0.13	0
		C.3.e	1.75	1	2.5	2

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		C.3	0.88	1	0	0
	C.4	C.4.a	0.5	0	0.25	0
		C.4.b	0.13	0	0.13	0
		C.4.c	3	2	4.25	3
		C.4.d	0	0	0	0
		C.4.e	2.5	2	1.5	1
		C.4.f	0.13	0	0.13	0
		C.4.g	0.25	0	0.13	0
		C.4	0.13	0	0.88	1
	C.5	C.5	1.25	1	1.13	1
		C.5.a	0	0	0.13	0
		C.5.b	2	1	1.5	1
		C.5.c	4.75	3	7.25	6
		C.5.d	2.5	2	1.5	1
		C.5.e	1	1	1	1
		C.5.f	1	1	1	1
		C.5.g	1.25	1	1.5	1
		C.5.h	0.13	0	0.63	0
	C.6	C.6.a	0	0	0.38	0
		C.6.b	4.5	3	3.38	3
		C.6.c	0.25	0	0.25	0
		C.6.d	0.75	1	0.25	0
		C.6.e	1.88	1	1.5	1
		C.6.f	0.75	1	1.63	1
		C.6.g	0	0	0	0
		C.6.h	0	0	0	0
		C.6	0.75	1	0.75	1

Percentages in table may not sum to 100% due to rounding.

As shown in Table 23, 75 of the 87 objectives received one or more hits in at least one panel; additionally, panelists coded items directly to nine goals and two standards. The three objectives having the greatest number of mean hits in both panels were all within Goal A.3, “Applications,” in the ACCUPLACER A, “Arithmetic,” set of objectives:

- A.3.a, “Rate problems including ratio and proportion”
- A.3.d, “Measurement problems”
- A.3.f, “Problems related to graphs and tables”

A majority of panelists across both panels identified ten items that were aligned to Objective A.3.a, eight items that were aligned to Objective A.3.d, and 12 items that were aligned to Objective A.3.f. The 30 items identified as aligned to all three of those objectives involve the application of one or more skills in a real-world context.

The following ACCUPLACER objectives received no hits in the set of 164 NAEP items:

- A.1.d, “Division of a whole number by a fraction or mixed number”
- A.1.e, “Division of a fraction/mixed number by a whole number, fraction or mixed number”
- A.1.h, “Recognition of equivalent fractions and mixed numbers”
- A.2.a, “Addition and subtraction of decimals”
- A.2.d, “Division of a decimal by a whole number other than a power of ten or by a decimal”
- A.2.f, “Division of a decimal by a power of ten”
- A.2.k, “A number when a percent of it is known”
- A.2.l, “Squares and square roots (decimals)”
- B.2.h, “Factoring polynomials that are not quadratics”
- C.4.d, “Determinants”
- C.6.g, “Trigonometric functions of two angles”
- C.6.h, “Inverse trigonometric functions”

The comments recorded by the panelists expressed several observations. Panelists remarked that, due to the nature of the ACCUPLACER objectives, many items were uncodable. As stated previously, many of the ACCUPLACER objectives are written to assess skills and concepts that are not necessarily at the grade 12 level, whereas the NAEP items are written at the grade 12 level. In addition, few ACCUPLACER objectives address the concepts of measurement, geometry, and data analysis and probability. For other items that were codable, panelists remarked that there were often parts of those items not addressed by the objective. The intent of the ACCUPLACER objectives, as inferred by the panelists, is primarily to assess discrete skills, whereas the NAEP items are frequently written at a higher level of cognitive demand and involve real-world context.

Table 24 displays the summary of alignment levels on three of the four content focus criteria for the alignment study: categorical concurrence, range of knowledge, and balance of representation. The ACCUPLACER framework was not able to be coded for depth of knowledge. The values in the table are intended to be descriptive only. For comparison purposes, asterisks are used to denote values considered “Weak” or “No” according to the typical WAT threshold values.

Table 24. Summary of Attainment of Acceptable Alignment Level on Three Content Focus Criteria as Rated by Eight Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 164[16]

	Alignment Criteria					
	Categorical Concurrence (mean hits)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
A – Arithmetic	50	42.25	39**	42*	0.59**	0.57**
B – Elementary algebra	43.88	35.12	62	65	0.65*	0.67*
C – College level math	47.62	48.5	52	55	0.73	0.73

One asterisk (*) indicates that the standard would weakly meet the alignment criterion according to the typical WAT threshold values. Two asterisks (**) indicate that the standard would not meet the alignment criterion according to the typical WAT threshold values.

Of the 164 NAEP assessment items analyzed, not all were found to match or "hit" objectives. As previously noted, there were a number of items found by panelists to be uncodable. However, using the typical WAT threshold value of six mean hits, categorical concurrence was met for ACCUPLACER A, "Arithmetic," with 42.25 (Panel 2) and 50 (Panel 1) mean hits to the standard. Categorical concurrence was also met for ACCUPLACER B, "Elementary algebra," with 35.12 (Panel 2) and 43.88 (Panel 1) mean hits to the standard, and ACCUPLACER C, "College level math," with 47.62 (Panel 1) and 48.5 (Panel 2) mean hits to the standard.

The range of knowledge criterion was met above the 50% criterion for "Elementary algebra" and for "College level math." Range of knowledge for "College level math" was greater than the baseline alignment of the ACCUPLACER short-version to the ACCUPLACER framework. Range of knowledge was not met or was met weakly for "Arithmetic." Most of the objectives within "Arithmetic" address skills that fall below the expectations of the typical grade 12 student. Therefore, although NAEP items were found to align to arithmetic objectives, there were many arithmetic objectives that did not receive hits from the grade 12 NAEP items; this is likely because arithmetic content tends to be targeted below the grade 12 level.

The balance of representation criterion is most relevant in this sub-study for "Arithmetic" and "Elementary algebra." Although a range of objectives received hits, the typical WAT threshold of 0.7 was not met for these two standards, with 0.57 and 0.67 mean hits for the two panels. This indicates that the items were clustered around only a few objectives in each goal. In this case, the items were clustered around Objectives A.3.a, A.3.d, A.3.f, A.3.g, B.2.a, B.2.d, and B.3.e.

As described earlier, the ACCUPLACER framework could not be coded for DOK. Table 25 shows the range of depth of knowledge of the items aligned to each ACCUPLACER standard.

[16] The percentages in this table indicate the distribution of total hits. It should be noted that, as shown in Table 21, 15% and 24% of the adjusted total hits for NAEP items were determined by panelists to be uncodable to any objective.

Table 25. Range of Depth of Knowledge of NAEP Items Aligned to the ACCUPLACER Framework

Assessment items = 164

	Panel 1							Panel 2						
		DOK Level 1		DOK Level 2		DOK Level 3			DOK Level 1		DOK Level 2		DOK Level 3	
Standards	**Mean Hits**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**
A	50	11.75	25	35.63	70	2.63	5	42.25	7.88	19	31.63	75	2.75	7
B	43.88	15.00	37	25.88	56	3.00	8	35.13	14.88	42	17.50	50	2.75	7
C	47.62	14.75	31	30.13	63	2.75	6	48.50	12.50	26	30.88	64	5.13	11
Total	141.5	41.50	37	91.63	59	8.38	4	125.88	35.25	34	80.00	59	10.63	7

As shown in Table 25, the NAEP items were found to be at DOK Levels 1–3, with most of the items at DOK Level 2. The items aligned for ACCUPLACER A, “Arithmetic,” were mostly at DOK Level 2 (70% in Panel 1 and 75% in Panel 2). Items aligned to ACCUPLACER B, “Elementary algebra,” were mostly at DOK Level 2 (56% in Panel 1 and 50% in Panel 2). Items aligned to ACCUPLACER C, “College level math,” were mostly at DOK Level 2 (63% in Panel 1 and 64% in Panel 2). In comparison to the baseline alignment of the ACCUPLACER short-version to the ACCUPLACER framework, NAEP had a wider range of depth of knowledge, with items at all three DOK levels aligned to all three standards. NAEP items aligned to “Arithmetic” and “Elementary algebra” had a greater percentage of Level 2 items than did ACCUPLACER items aligned to these standards. For “College level math,” the distribution of Level 1 and Level 2 items for NAEP items and ACCUPLACER short-version items was more similar.

IV. Panelists' Evaluations of the Process

This section details the findings from responses to training and process evaluation questionnaires that the eight panelists from each of two panels (a total of 16 panelists) completed at the end of each day of participation. WestEd staff administered these questionnaires to determine what factors, if any, might impede consistent and reliable alignment coding within and across panels. WestEd staff compiled and reviewed the responses daily to identify necessary refinements to study logistics and/or needs for additional panelist training and to inform discussions with facilitators as necessary to ensure ongoing accurate application of the study protocol. Each questionnaire asked panelists to indicate her/his participant number, content area, and group number. In addition, questionnaires had 14 (Day 1), 8 (Day 2, Day 3, and Day 4), and 17 (Day 5) substantive questions. This analysis compares panelist responses across the two panels; in addition, for questions that were repeated across multiple questionnaires, responses are compared across days. Full verbatim responses to all questionnaires are included in Appendix I.

Day 1 Training and Process Evaluation

Following the first day of the study, panelists were administered a questionnaire that solicited feedback on the training for assigning DOK values to objectives and on the first day's alignment activities. Table 26 shows results for selected-response questions 5–9, 12, and 13, by panel. Numbers in bold font represent the highest number of responses for each question, by panel.

Table 26. Panelist Responses to Day 1 Training and Process Evaluation Questionnaire

	Panel 1 (n=8)				**Panel 2** (n=8)			
How well did the training…	**Not Well**	**Some-what**	**Ade-quately**	**Very Well**	**Not Well**	**Some-what**	**Ade-quately**	**Very Well**
Q5. explain the purpose of the study?	0	0	**4**	**4**	0	0	2	**6**
Q6. introduce NAEP/ ACCUPLACER?	0	0	**5**	3	0	0	3	**5**
Q7. prepare you to under-stand DOK levels?	0	0	**6**	2	0	0	3	**5**
Q8. prepare you for the consensus process?	0	1	**6**	1	0	0	2	**6**
Q9. prepare you to use the WAT system?	0	0	2	**6**	0	0	2	**6**
How comfortable do you feel…	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**
Q12. assigning DOK levels to objectives?	0	**4**	2	2	0	0	**4**	**4**

	Panel 1 (n=8)			Panel 2 (n=8)		
How well did your facilitator…	**Not Well**	**Moderately Well**	**Very Well**	**Not Well**	**Moderately Well**	**Very Well**
Q13. facilitate today's consensus process?	0	**4**	**4**	0	1	**7**

As shown in Table 26, all panelists across the two panels reported that the introductory session either adequately or very well explained the purposes of the study, introduced the NAEP and ACCUPLACER assessments, and prepared them for understanding definitions of DOK levels. All but one panelist felt the training prepared them adequately or very well for the discussion process that led to agreement on DOK levels for NAEP objectives across the two panels; the panelist who reported that the training prepared him/her somewhat well would have preferred the panel to have formed some shared meanings for terminology encountered in the NAEP framework. When asked their level of comfort with assigning DOK levels to objectives, the majority of the panelists across the two panels (75%, or 12 panelists) felt comfortable or very comfortable doing so, while the remaining 25% (4) felt somewhat comfortable. All four panelists who reported feeling somewhat comfortable with assigning DOK levels were participants from Panel 1. One panelist suggested that it might be beneficial to have a pilot coding session on selected NAEP objectives. Another panelist felt that some objectives were difficult to code because the DOK levels were not defined clearly enough or did not allow for distinctions of complexity across items. All the panelists reported that the discussions that led to agreement on DOK levels for NAEP objectives across the two panels were facilitated either moderately or very well. Regarding use of the WAT, none of the panelists had used the WAT prior to this study, and all of the panelists reported feeling either adequately or very well prepared to use the WAT system.

This questionnaire provided opportunities for panelists to indicate aspects of the day's alignment tasks that went well or not well, to make suggestions for improving the alignment activities, and to raise concerns or questions about the alignment process. Recommendations for improving the training and alignment process included providing more examples of difficult key concept items that had been coded, providing time for a pilot coding prior to the actual coding, more focus on the discussion process, clearer definitions of the DOK levels for some objectives, more discussion about the verbs used in the items, and time to form shared meanings for terms used in the NAEP framework. When asked what aspects of the day went particularly well, responses included the group discussions, the adjudication process, the small group DOK coding, and the mixture of panelists in the groups.

WestEd staff used this feedback to evaluate whether the alignment process could continue on Day 2 as scheduled; they determined that all panelists were sufficiently trained in the Day 1 assignment of DOK levels to objectives, and felt confident moving into the Day 2 activities. WestEd staff monitored both panels to ensure that all panelists felt comfortable completing, and were able to complete, the remaining alignment activities in accordance with the training.

Day 2 Training and Process Evaluation

On the second day of the study, panelists were trained in assigning DOK values to items and in determining alignments to objectives; they then mapped NAEP items to the NAEP framework and reviewed the ACCUPLACER framework. At the end of the day, they were administered a training and process evaluation questionnaire that solicited feedback on the training and alignment activities. Table 27 shows results for selected-response questions 4, 5, and 6, by panel. Numbers in bold font represent the highest number of responses for each question, by panel.

Table 27. Panelist Responses to Day 2 Training and Process Evaluation Questionnaire

	Panel 1 (n=8)				**Panel 2 (n=8)**			
How well did the training…	**Not Well**	**Some-what**	**Ade-quately**	**Very Well**	**Not Well**	**Some-what**	**Ade-quately**	**Very Well**
Q4. prepare you to assign DOK levels to test items?	0	0	0	8	0	0	3	5
Q5. prepare you for the alignment (coding) process?	0	0	1	7	0	0	3	5
How well did your facilitator…	**Not Well**	**Moderately Well**		**Very Well**	**Not Well**	**Moderately Well**		**Very Well**
Q6. facilitate today's consensus process?	0	1		7	0	0		8

Panelists reported feeling similarly prepared by the training to both assign DOK values to items and perform the alignment (coding) process; across the two panels, all of the panelists reported feeling either adequately or very well trained for both of these activities. One panelist reported that her/his facilitator conducted the day's within-panel discussions regarding alignment codes moderately well.[17] The remaining 15 panelists indicated that the facilitators conducted the discussions very well.

A recommendation for improving the alignment process was to provide smaller blocks of items to align in one sitting. When asked what activities went particularly well, three panelists commented on the value of consensus discussions, with additional comments relating to the alignment process. Panelists did not report any activities they felt unprepared for or any items about which they wanted more information.

Day 3 Process Evaluation

The third day of the study comprised mapping of both ACCUPLACER and NAEP items to the ACCUPLACER framework. At the end of the day, panelists were administered a process evaluation questionnaire that solicited feedback on these alignment activities. Table 28 shows results for selected-response questions 4, 5, and 6, by panel. Numbers in bold font represent the highest number of responses for each question, by panel. One Panel 1 panelist did not complete her/his questionnaire; therefore, only seven responses are reported for Panel 1.

[17] The within-panel discussions were referred to in the questionnaire as the "consensus process," although it was understood that true consensus was neither a requirement nor a goal, per the study design document.

Table 28. Panelist Responses to Day 3 Process Evaluation Questionnaire

	Panel 1 (n=7)				**Panel 2** (n=8)			
How comfortable do you feel…	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**
Q4. assigning DOK levels to test items?	0	0	0	**7**	0	0	**4**	**4**
Q5. aligning test items to objectives?	1	2	**3**	1	0	2	**6**	0
How well did your facilitator…	**Not Well**	**Moderately Well**	**Very Well**		**Not Well**	**Moderately Well**	**Very Well**	
Q6. facilitate today's consensus process?	0	1	**6**		0	1	**7**	

Across both panels, participants generally felt more comfortable assigning DOK levels to test items than they felt aligning test items to objectives. All panelists felt either comfortable or very comfortable assigning DOK levels to test items. However, while 67% (10) of the panelists felt comfortable or very comfortable aligning test items to objectives, the remaining 25% (4) of the panelists felt somewhat comfortable and 6% (1) felt uncomfortable aligning test items to objectives. The one panelist from Panel 1 who reported being uncomfortable aligning test items to objectives did not have any specific suggestions for improving the alignment process. However, this panelist felt the alignment process was tedious and that there were too many items to review in the amount of time allotted. A panelist from Panel 2 who was somewhat comfortable with the alignment process also felt that additional time would be beneficial and would enable panelists to take longer breaks to refresh.

Of the two panelists from the first panel who were somewhat uncomfortable aligning test items to the objectives, one offered no suggestions or responses to any of the open-ended questions. The second panelist felt the need for more direction from the facilitators on how to align NAEP to ACCUPLACER, since the NAEP assessment was not designed to match the ACCUPLACER framework; this suggestion was also mentioned by one of the panelists from Panel 2.

As on the first two days of the study, all panelists felt that the facilitator conducted the day's within-panel discussions either moderately or very well.

The questionnaire also asked panelists to provide recommendations for improving the alignment process, to record requests for more information, and to specify activities that they felt unprepared for. Panelists felt that additional definitions of terms (e.g., graphics), more time for the alignments, and more directions on alignment would improve the alignment process. When asked what activities went particularly well, one panelist expressed that her/his group facilitator kept the panelists focused, while four other panelists mentioned that the group discussions were particularly positive. When asked what could have been done differently to improve the day's activities, suggestions included fewer items and breaking the coding into smaller segments.

Day 4 Process Evaluation

The Day 4 alignment activities included completion of the coding of NAEP items to the ACCUPLACER framework and beginning coding of the ACCUPLACER items to the NAEP framework. At the end of the day, panelists were administered a process evaluation questionnaire that solicited feedback on these alignment activities. Table 29 shows results for selected-response questions 4, 5, and 6, by panel. Numbers in bold font represent the highest number of responses for each question, by panel.

Table 29. Panelist Responses to Day 4 Process Evaluation Questionnaire

	Panel 1 (n=8)				**Panel 2** (n=8)			
How comfortable do you feel…	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**
Q4. assigning DOK levels to test items?	0	1	1	**6**	0	1	2	**5**
Q5. aligning test items to objectives?	0	**3**	2	**3**	0	1	**5**	2
How well did your facilitator…	**Not Well**	**Moderately Well**	**Very Well**		**Not Well**	**Moderately Well**	**Very Well**	
Q6. facilitate today's consensus process?	0	0	**8**		0	1	**7**	

All but two panelists, one from each panel, reported feeling comfortable or very comfortable with assigning DOK values to items (the remaining two panelists felt somewhat comfortable, although neither one made suggestions for improving the alignment process or noted feeling unprepared for this activity). In comparison, panelists reported less comfort with coding ACCUPLACER items to the NAEP framework. Twenty-five percent (4) of the panelists across the two panels reported feeling somewhat comfortable with this task, with the remaining 75% (12) of the panelists feeling either comfortable or very comfortable. Again, the majority (94%, or 15) of panelists reported that the facilitators conducted the day's within-panel discussions very well, with one panelist rating the panel's facilitation as conducted moderately well.

Panelists were asked to provide recommendations for improving the alignment process, to record requests for more information, and to specify activities they felt unprepared for. Four panelists expressed concern over the number of items to code, including that the workload made them feel rushed. Extended breaks for facilitator discussions are a feature of the study design and were described during the opening training, but three panelists expressed concern about the two-hour afternoon break on Day 4 while the facilitators compared results for the cross-panel adjudication process. Two of these panelists suggested allowing panels to move on to the next coding step as facilitators conducted cross-panel adjudication. Overall, however, responses were largely positive, with panelists again expressing appreciation for opportunities to discuss and adjudicate alignment decisions. One panelist described the day as "a little pushed, but extremely productive."

Day 5 End-of-Study Evaluation

On the final day of the study, panelists responded to additional questions about the study logistics, the effectiveness of their panels and facilitators, and the alignment process. Responses to this questionnaire were used by WestEd staff as a final opportunity to identify and address any issues with panelist alignment codes and to identify deficiencies in training or workshop logistics that could be addressed for future alignment studies. Table 30 shows results for selected-response questions 4–11 and 15, by panel. Numbers in bold font represent the highest number of responses for each question, by panel.

Table 30. Panelist Responses to End-of-Study Evaluation Questionnaire

	Panel 1 (n=8)				**Panel 2** (n=8)			
How well did Monday's training prepare you…	**Not Well**	**Some-what**	**Ade-quately**	**Very Well**	**Not Well**	**Some-what**	**Ade-quately**	**Very Well**
Q4. for understanding DOK levels?	0	0	**7**	1	0	0	**5**	3
Q7. for the consensus process?	0	2	**3**	**3**	0	1*	**5***	1*
Q8. for the alignment (coding) process?	0	0	**6**	2	0	1	**4**	3
How comfortable did you feel…	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**
Q5. assigning DOK levels to objectives?	0	1	6	1	0	0	**4**	**4**
Q6. assigning DOK levels to test items?	0	0	4**	**4**	0	0	**4**	**4**
How useful was/were…	**Not Useful**	**Some-what Useful**	**Ade-quately Useful**	**Very Useful**	**Not Useful**	**Some-what Useful**	**Ade-quately Useful**	**Very Useful**
Q9. information provided prior to the study?	0	**6**	2	0	2	6	0	0
Q10. on-site training and coding materials?	0	0	2	6	1	0	5	2
How qualified was your panel…	**Not Quali-fied**	**Some-what Quali-fied**	**Ade-quately Quali-fied**	**Very Quali-fied**	**Not Quali-fied**	**Some-what Quali-fied**	**Ade-quately Quali-fied**	**Very Quali-fied**
Q11. to conduct this type of alignment?	0	0	4**	4	0	0	0	**8**

	Panel 1 (n=8)				Panel 2 (n=8)			
How easy was it…	**Not Easy**	**Some-what Easy**	**Adeq-uately Easy**	**Very Easy**	**Not Easy**	**Some-what Easy**	**Adeq-uately Easy**	**Very Easy**
Q15. to use the WAT for the alignment process?	1	1	3	3	0	0	4	4

*n=7 for this question.
**If a panelist's response fell between two categories, it was counted in the lower of the two categories.

At the end of the week, as evidenced on the end-of-study questionnaire, all panelists across the two panels reported feeling adequately or very well trained by the Day 1 training in understanding DOK levels, and all but one panelist felt that the Day 1 training adequately or very well prepared them for the alignment coding process. The remaining panelist felt somewhat well trained by the Day 1 training. When asked about their comfort level with assigning DOK levels to objectives and items, all panelists reported being comfortable or very comfortable, except for one panelist in Panel 1 who reported being somewhat comfortable assigning DOK levels to objectives.

The majority (88%, or 14) of the panelists indicated that the information they received in advance of the study was adequately or somewhat useful; two panelists indicated that the pre-study materials were not useful, although they did not elaborate on this response. In addition, all but one of the panelists felt that the training and coding materials they received during the alignment institute were either very or adequately useful, with the remaining panelist indicating that the materials were not useful. All the panelists reported that their panels were either adequately qualified or very qualified to conduct the alignment activities; of those panelists who provided comments about the composition of the panel (12 panelists), all but two felt the panel compositions were effective for this type of study. One of these two panelists, from Panel 1, suggested that experience with standards, assessments, and alignment would support the evaluation of standards; the second of these two panelists, from Panel 2, recommended more culturally diverse panels.

Panelists from both panels reported that the adjudication procedures implemented by their facilitators were effective. Regarding the alignment process and tools, overall, panelists responded positively to the WAT, although some commented that the system might work better for panelists if it were in a spreadsheet environment, included better navigational tools, or did not have technical issues such as screen jumping.

Panelists reported that the alignment criteria were useful in capturing what the panelists found to be important aspects of each assessment item or objective. In response to questions about how well the process captured content similarities and differences between the assessments, panelists reported that the process captured them well, and provided many examples of their impressions of those similarities and differences, although some panelists indicated that there were more differences than similarities. A few panelists also expressed that there were challenges associated with aligning an assessment to a framework to which it was not developed, and several panelists commented that NAEP was more concept- and process-focused, while ACCUPLACER focused on skills and content.

The final evaluation survey also included a question about the assessments themselves, and panelists reported similarities and differences related to item types, skills measured, DOK levels, frameworks, and content assessed in the assessments.

When asked about the facilities for the alignment workshop, panelists felt that they were suitable. Table 31 shows panelist responses to this question, by panel. Numbers in bold font represent the highest number of responses for each question, by panel. One panelist from Panel 1 did not complete this questionnaire; therefore, only seven responses are reported for Panel 1.

Table 31. Panelist Responses Regarding Adequacy of Facilities

	Panel 1 (n=7)				**Panel 2** (n=8)			
How suitable were the facilities for this workshop…	**Not Suitable**	**Some-what Suitable**	**Ade-quately Suitable**	**Very Suitable**	**Not Suitable**	**Some-what Suitable**	**Ade-quately Suitable**	**Very Suitable**
Meeting rooms	0	1	1	**5**	0	1	3	**4**
Computers and equipment	0	1	1	**5**	0	1	2	**5**
Meals and breaks	0	0	1	**6**	0	**4**	1	3
Sleeping rooms	0	0	0	**7**	0	0	1	**7**

Across both panels, 87% (13) of the panelists who responded felt that the meeting rooms were very suitable or adequately suitable, while the remaining 13% (2) of the responding panelists found the meeting rooms somewhat suitable. Neither had specific comments about the rooms, although one panelist from Panel 2 noted that the conference area was only accessible via the elevators and that s/he would have preferred to use the stairs. Similarly, 87% (13) of the panelists who responded felt that the computers and equipment were either very suitable or adequately suitable for this type of alignment workshop, while the remaining 13% (2) of the respondents felt that the computers and equipment were somewhat suitable. Regarding the meals and breaks, 73% (11) of the panelists felt that they were either very suitable or adequately suitable for this type of alignment workshop, while the remaining 27% (4) responded that the meals and breaks were somewhat suitable. Of the four panelists who responded that the meals and breaks were somewhat suitable, one panelist felt the work was mentally draining and that more breaks were needed during the day. Another analyst felt that more time should have been allotted for the mathematics group than for the reading group, as the mathematics panelists found themselves working through lunch at times or either starting early or working late in order to complete all alignment tasks. This panelist felt that, for future studies, consideration should be given not only to the number of assessment items, but also to the number of objectives from which panelists had to choose. All the panelists who responded felt the sleeping rooms were either very or adequately suitable.

Overall, panelists enjoyed the opportunity to participate in this study and felt that the meetings were well planned and well organized. One panelist suggested that, since all panelists had computers, debrief surveys be administered online, which would eliminate the need to handwrite comments. Two panelists mentioned the need for more time to accomplish their tasks; although they found the work enjoyable, they felt mentally fatigued at the end of the day. There was a

perceived inequity between the amount of work done by the mathematics group in comparison to the reading group, due to the number of test items in each pool and the number of objectives in each framework. A panelist from Panel 1 felt that his/her group did not fully understand the distinction between difficulty and complexity, which led to prolonged discussion and some differences in coding. Although WestEd staff inquired in advance about panelists' dietary restrictions, one panelist reported wanting fewer carbohydrates and more protein. One panelist found the meeting rooms too cool and wished that they had windows.

V. Summary and Conclusions

Section III reported various indices of alignment for each sub-study individually. This section compares the results of the sub-studies in terms of the overlap of the content alignment of each test, including a summary of alignment of each assessment vis-à-vis the four criteria of the study. The section ends with overall conclusions regarding the alignment of the NAEP and ACCUPLACER mathematics assessments.

Summary of Overlap of Content Alignment

Table 32 shows the overlap of content alignment of each assessment to its own and the other's framework in terms of the percentages of total hits.

Table 32. Summary of the Overlap of Content Alignment between NAEP and ACCUPLACER Items and the NAEP and ACCUPLACER Frameworks at the Standard Level

	NAEP Items (42 items)		**ACCUPLACER Items** (105 items)	
	Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
NAEP Framework	**% of Total Hits**		**% of Total Hits**	
1 – Number properties and operations	19%	19%	38%	39%
2 – Measurement	18%	15%	3%	4%
3 – Geometry	16%	17%	1%	0%
4 – Data analysis, statistics, and probability	10%	12%	2%	2%
5 – Algebra	37%	37%	56%	54%
	NAEP Items (164 items)[18]		**ACCUPLACER Items** (45 items)	
ACCUPLACER Framework	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
A – Arithmetic	35%	34%	33%	35%
B – Elementary algebra	31%	28%	39%	36%
C – College level math	34%	39%	27%	29%

Percentages in table may not sum to 100% due to rounding.

NAEP items were found to assess all NAEP standards (Standard 1, "Number properties and operations"; Standard 2, "Measurement"; Standard 3, "Geometry"; Standard 4, "Data analysis, statistics, and probability"; and Standard 5, "Algebra") and all of the three ACCUPLACER standards (ACCUPLACER A, "Arithmetic"; ACCUPLACER B, "Elementary algebra"; and ACCUPLACER C, "College level math"). ACCUPLACER items were found to assess all of the ACCUPLACER standards ("Arithmetic," "Elementary algebra," and "College level math") and, to varying degrees, four NAEP standards: "Number properties and operations," "Measurement," "Data analysis, statistics, and probability," and "Algebra." However, coverage of "Measurement" and "Data analysis, statistics, and probability" was limited, with less than 5% of

[18] The percentages in this table indicate the distribution of total hits. It should be noted that, as shown in Table 21, 15% and 24% of the adjusted total hits for NAEP items were determined by panelists to be uncodable to any objective.

overall hits. ACCUPLACER items also had very limited coverage of “Geometry” in Panel 1 (1%) and nearly no coverage of “Geometry” in Panel 2 (0%).

With regard to alignment to the NAEP framework, “Algebra” had the highest percentage of overall hits from both ACCUPLACER items (54% and 56% of total hits) and NAEP items (37% of total hits). The remaining NAEP short-version sample set item alignments were distributed more evenly among the four remaining standards, with 15–19% to “Number properties, and operations,” “Measurement,” and “Geometry,” and 10% and 12% to “Data analysis, statistics, and probability.” After “Algebra,” most of the remaining hits for ACCUPLACER items (38% and 39%) were distributed to “Number properties and operations,” while “Measurement” and “Data analysis, statistics, and probability” each received less than 5% of the overall alignments. ACCUPLACER had little to no coverage of “Geometry.”

In relation to the ACCUPLACER framework, both assessments had somewhat similar distributions of item alignments among the three ACCUPLACER standards. However, it should be noted that 15% and 24% of adjusted total hits for the NAEP items were uncodable; 20 items were found to be uncodable by the majority of panelists in each panel (5 or more of 8) and not aligned to any ACCUPLACER objective. The uncodable NAEP items tended to be those determined by panelists to assess “Measurement,” “Geometry,” and “Data analysis, statistics, and probability.” The NAEP items that did align had 34% and 35% of hits to “Arithmetic,” and 34% and 39% to “College level math.” “Elementary algebra” received the fewest hits, with 28% and 31% of the NAEP alignments. In the ACCUPLACER assessment, “Elementary algebra” received the highest percentage (36% and 39%) of item alignments, followed closely by “Arithmetic” (33% and 35%). “College level math,” which had the highest percentage of hits in the NAEP assessment, received the lowest percentage of ACCUPLACER item alignments, with 27% and 29% of total hits. Despite these differences in emphasis, however, considering only codable items, the percentages of alignments to each ACCUPLACER standard were relatively evenly distributed in both assessments and similar in distribution across assessments.

Overall, the NAEP items covered all of the ACCUPLACER objectives with a slightly stronger emphasis on “Arithmetic” and “College level math.” ACCUPLACER items covered all of the ACCUPLACER objectives, with “Arithmetic” and “Elementary algebra” receiving slightly greater emphasis.

Overlap in content alignment to the NAEP framework can also be examined at the more finely grained objective level. Table 33 shows the overlap of alignment of each assessment to the NAEP framework in terms of the percentages of total hits.

Table 33. Summary of the Overlap of Content Alignment between NAEP and ACCUPLACER Items and the NAEP Framework at the Objective Level

			NAEP Items (42 items)		**ACCUPLACER Items** (105 items)	
			Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
Standards	**Goals**	**Objectives**	**% of Total Hits**		**% of Total Hits**	
1	1.1	1.1.d	1	2	2	2
		1.1.f	2	2	0	0
		1.1.g	0	0	1	1
		1.1.i	0	0	4	4
	1.2	1.2.b	1	1	0	0
		1.2.c	0	2	0	0
		1.2.d	0	0	1	1
		1.2	0	0	0	0
	1.3	1.3.a	0	0	1	1
		1.3.b	5	3	18	19
		1.3.c	2	2	1	0
		1.3.d	0	0	0	0
		1.3.f	2	1	5	5
		1.3	0	0	0	0
	1.4	1.4.c	0	0	2	2
		1.4.d	2	2	2	3
		1.4	0	0	1	0
	1.5	1.5.c	1	2	0	0
		1.5.d	0	0	0	0
		1.5.e	0	0	0	0
		1.5.f	0	0	0	0
	1.6	1.6.a	1	0	0	0
		1.6.b	1	0	0	0
2	2.1	2.1.b	0	0	0	0
		2.1.c	1	0	0	0
		2.1.d	1	0	0	0
		2.1.f	1	1	0	1
		2.1.h	0	0	0	0
		2.1.i	3	2	0	0
		2.1	0	0	0	0
	2.2	2.2.a	2	2	0	0

Standards	Goals	Objectives	NAEP Items (42 items)		ACCUPLACER Items (105 items)	
			Panel 1	Panel 2	Panel 1	Panel 2
			% of Total Hits		% of Total Hits	
		2.2.b	3	1	0	1
		2.2.d	0	0	0	0
		2.2.e	0	1	0	0
		2.2.f	5	6	0	0
	2.3	2.3.a	0	0	0	0
		2.3.b	0	0	0	0
		2.3.c	1	2	2	2
		2.3.d	0	0	0	0
		2.3.e	0	0	0	0
		2.3.f	0	0	0	0
		2.3.g	0	0	0	0
		2.3	0	0	0	0
3	3.1	3.1.c	0	0	0	0
		3.1.d	1	1	0	0
		3.1.e	0	0	0	0
		3.1.f	0	0	0	0
	3.2	3.2.a	0	0	0	0
		3.2.b	0	0	0	0
		3.2.c	1	1	0	0
		3.2.d	1	2	0	0
		3.2.e	0	0	0	0
		3.2.g	0	0	0	0
	3.3	3.3.b	3	1	0	0
		3.3.c	0	0	0	0
		3.3.d	1	2	0	0
		3.3.e	0	0	0	0
		3.3.f	2	2	0	0
		3.3.g	0	0	0	0
		3.3.h	1	1	0	0
	3.4	3.4.a	2	2	1	0
		3.4.b	0	0	0	0
		3.4.c	0	2	0	0
		3.4.d	0	0	0	0

Standards	Goals	Objectives	NAEP Items (42 items)		ACCUPLACER Items (105 items)	
			Panel 1	Panel 2	Panel 1	Panel 2
			% of Total Hits		% of Total Hits	
		3.4.e	0	0	0	0
		3.4.f	0	0	0	0
		3.4.g	0	0	0	0
		3.4.h	2	2	0	0
	3.5	3.5.a	0	0	0	0
		3.5.b	0	0	0	0
		3.5.c	0	0	0	0
		3.5.d	0	0	0	0
		3.5.e	0	0	0	0
4	4.1	4.1.a	2	4	1	1
		4.1.b	1	0	0	0
		4.1.c	0	0	0	0
		4.1.d	3	0	0	0
		4.1.e	0	0	0	0
		4.1.f	0	0	0	0
	4.2	4.2.a	0	1	1	1
		4.2.b	0	0	0	0
		4.2.c	0	0	0	0
		4.2.d	1	0	0	0
		4.2.e	0	0	0	0
		4.2.f	0	0	0	0
		4.2.g	0	0	0	0
	4.3	4.3.a	0	0	0	0
		4.3.b	0	0	0	0
		4.3.c	1	2	0	0
		4.3.d	0	0	0	0
		4.3.e	0	0	0	0
	4.4	4.4.a	0	0	0	0
		4.4.b	1	0	0	0
		4.4.c	0	0	0	0
		4.4.d	0	0	0	0
		4.4.e	0	0	1	1
		4.4.h	1	2	0	0

Standards	Goals	Objectives	NAEP Items (42 items)		ACCUPLACER Items (105 items)	
			Panel 1	Panel 2	Panel 1	Panel 2
			% of Total Hits		% of Total Hits	
		4.4.i	0	0	0	0
		4.4.j	1	0	0	0
		4.4.k	0	0	0	0
	4.5	4.5.a	0	2	0	0
		4.5.b	0	0	0	0
		4.5.c	0	0	0	0
		4.5.d	0	0	0	0
		4.5.e	0	0	0	0
5	5.1	5.1.a	3	4	0	1
		5.1.b	4	2	0	0
		5.1.e	4	8	5	3
		5.1.g	0	0	0	0
		5.1.h	0	0	1	1
		5.1.i	1	0	1	1
		5.1.j	0	0	1	1
		5.1	0	0	0	0
	5.2	5.2.a	2	2	4	3
		5.2.b	0	0	0	0
		5.2.d	0	0	1	2
		5.2.e	2	1	0	0
		5.2.f	3	2	0	0
		5.2.g	0	0	0	0
		5.2.h	0	0	0	1
		5.2	1	0	0	0
	5.3	5.3.b	1	0	2	3
		5.3.c	2	2	15	19
		5.3.d	2	3	4	1
		5.3.e	3	2	3	3
		5.3.f	1	2	1	1
		5.3.g	0	0	0	0
		5.3.h	0	0	0	0
		5.3	0	0	0	0
	5.4	5.4.a	2	2	9	9

Standards	Goals	Objectives	NAEP Items (42 items)		ACCUPLACER Items (105 items)	
			Panel 1	Panel 2	Panel 1	Panel 2
			% of Total Hits		% of Total Hits	
		5.4.c	0	1	2	2
		5.4.d	0	0	4	3
		5.4.e	0	0	0	0
		5.4.f	0	0	1	1
		5.4.g	0	0	0	0
		5.4	0	0	0	0
	5.5	5.5.a	3	3	0	0
		5.5.b	1	2	0	0
		5.5.c	0	0	0	0

Percentages in table may not sum to 100% due to rounding.

As shown in Table 33, the NAEP items were much more evenly distributed across the NAEP objectives than were the ACCUPLACER items, with broader overall coverage of the NAEP objectives (75 objectives hit, compared to 56 for ACCUPLACER). Most of the hits in the NAEP assessment ranged between 1% and 4% of total hits per objective. Only two NAEP objectives received more than 4% of the total hits: 1.3.b ("Perform arithmetic operations with real numbers, including common irrational numbers") and 5.1.e ("Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or *trigonometric functions from tables, graphs, or equations"). In contrast, the majority of hits (two-thirds) from ACCUPLACER items were to just two NAEP objectives: 1.3.b (18% and 19% of total hits) and 5.3.c ("Perform basic operations, using appropriate tools, on algebraic expressions including polynomial and rational expressions") (15% and 19% of total hits). Three other objectives received over 4% of the ACCUPLACER hits: 5.4.a ("Solve linear, rational, or quadratic equations or inequalities, including those involving absolute value"); 5.1.e; and 1.3.f ("Solve application problems involving numbers, including rational and common irrationals").

Overall, the 42 NAEP items provided considerably broader coverage of the NAEP framework than did the 105 ACCUPLACER items. For instance, in both panels, NAEP items were aligned to 17 NAEP objectives that each received less than 1% of total hits from ACCUPLACER items: eight objectives in "Geometry," most of those to objectives for Goal 3.3 ("Relationships between geometric figures"); five objectives for "Algebra"; three objectives for "Measurement"; and two objectives each for both "Data analysis, statistics, and probability" and "Number properties and operations."

The following NAEP objectives had hits from the NAEP items in the short version (42 items) but less than 1% of hits from the full form of ACCUPLACER (105 items):

- 1.1.f, "Represent or interpret expressions involving very large or very small numbers in scientific notation"

- 1.2.b, “Identify situations where estimation is appropriate, determine the needed degree of accuracy, and analyze* the effect of the estimation method on the accuracy of results”
- 2.1.i, “Solve problems involving rates such as speed, density, population density, or flow rates”
- 2.2.a, “Recognize that geometric measurements (length, area, perimeter, and volume) depend on the choice of a unit, and apply such units in expressions, equations, and problem solutions”
- 2.2.f, “Construct or solve problems involving scale drawings”
- 3.1.d, “Draw or sketch from a written description plane figures and planar images of three-dimensional figures”
- 3.2.d, “Identify transformations, combinations, or subdivisions of shapes that preserve the area of two-dimensional figures or the volume of three-dimensional figures”
- 3.3.b, “Apply geometric properties and relationships to solve problems in two and three dimensions”
- 3.3.f, “Analyze properties or relationships of triangles, quadrilaterals, and other polygonal plane figures”
- 3.3.h, “Analyze properties of circles and the intersections of lines and circles (inscribed angles, central angles, tangents, secants, and chords)”
- 3.4.h, “*Represent situations and solve problems involving polar coordinates”
- 4.3.c, “*Draw inferences from samples, such as estimates of proportions in a population, estimates of population means, or decisions about differences in means for two ‘treatments’”
- 4.4.h, “Determine the probability of independent and dependent events”
- 5.1.b, “Express linear and exponential functions in recursive and explicit form given a table, verbal description, or some terms of a sequence”
- 5.2.e, “Make inferences or predictions using an algebraic model of a situation”
- 5.2.f, “Given a real-world situation, determine if a linear, quadratic, rational, exponential, logarithmic, or *trigonometric function fits the situation”
- 5.5.a, “Use algebraic properties to develop a valid mathematical argument”
- 5.5.c, “Explain the use of relational conjunctions (and, or) in algebraic arguments”

Conversely, ACCUPLACER items were mapped by both panels to ten NAEP objectives for which the NAEP short-version items had less than 1% of total hits from either panel. These included five objectives for “Number properties and operations,” four objectives for “Algebra,” and one objective for “Data analysis, statistics, and probability.” It is important to remember, however, that the NAEP items represented a relatively small subset (42) of the complete pool of 164 NAEP items, while the ACCUPLACER set included 105 items.

The following NAEP objectives had better coverage by the full ACCUPLACER form (105 items) than the NAEP short version of 42 items; these objectives had hits from ACCUPLACER items and less than 1% of hits from NAEP:

- 1.1.g, “Represent, interpret, or compare expressions or problem situations involving absolute values”
- 1.1.i, “Order or compare real numbers, including very large and very small real numbers”

- 1.2.d, “Estimate square or cube roots of numbers less than 1,000 between two whole numbers”
- 1.3.a, “Find integral or simple fractional powers of real numbers”
- 1.4.c, “Use proportions to solve problems (including rates of change)”
- 4.4.e, “Determine the number of ways an event can occur using tree diagrams, formulas for combinations and permutations, or other counting techniques”
- 5.1.h, “Recognize and analyze the general forms of linear, quadratic, rational, exponential, or *trigonometric functions”
- 5.1.j, “*Given a function, determine its inverse if it exists and explain the contextual meaning of the inverse for a given situation”
- 5.2.d, “Perform or interpret transformations on the graphs of linear, quadratic, exponential, and *trigonometric functions”
- 5.4.d, “Solve (symbolically or graphically) a system of equations or inequalities and recognize the relationship between the analytical solution and graphical solution”
- 5.4.f, “Solve an equation or formula involving several variables for one variable in terms of the others”

Differences in breadth and emphasis for the two assessments were evident across all standards and objectives. As indicated in the preceding lists, for example, each assessment had hits to five “Algebra” objectives, for which the other assessment had less than 1% total hits, reflecting differences in emphasis within that standard. Furthermore, the majority of the ACCUPLACER hits to “Algebra” (15% and 19% of total hits) were to one objective, 5.3.c (“Perform basic operations, using appropriate tools, on algebraic expressions including polynomial and rational expressions”), while the NAEP hits were much more evenly distributed across objectives, with only Objective 5.1.e (“Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or *trigonometric functions from tables, graphs, or equations”) receiving more than 4% of total hits. The ACCUPLACER items had very limited coverage of “Data analysis, statistics, and probability,” with only three objectives (4.1.a, 4.2.a, and 4.4.e) receiving 1% of total hits each. The NAEP hits were distributed to ten objectives for “Data analysis, statistics, and probability,” with Objective 4.1.a (“Read or interpret graphical or tabular representations of data”) receiving the highest percentage of hits. Compared to ACCUPLACER, the NAEP items also had much stronger coverage of objectives in “Geometry,” with hits to ten objectives and four of the five goals. Objectives for Goals 3.3 (“Relationships between geometric figures”) and 3.4 (“Position, direction, and coordinate geometry”) had the most NAEP hits. The ACCUPLACER items had almost no coverage of “Geometry,” with only Objective 3.4.a (“Solve problems involving the coordinate plane such as the distance between two points, the midpoint of a segment, or slopes of perpendicular or parallel lines”) receiving 1% of total hits from one panel only. Coverage of “Measurement” objectives was also notably less for the ACCUPLACER items. Only three objectives (2.1.f, 2.2.b, and 2.3.c) received hits from ACCUPLACER items. However, the NAEP items covered nine objectives and all three goals for “Measurement,” with Objective 2.2.f (“Construct or solve problems involving scale drawings”) receiving the highest percentage of hits.

In terms of the number of objectives receiving hits, the two tests had similar results for “Number properties and operations”: both the NAEP items and the ACCUPLACER items had hits to 11 objectives. As with “Algebra,” however, ACCUPLACER hits to “Number properties and operations” were concentrated on a single objective, 1.3.b, which received 18% and 19% of total

hits. Of the other “Number properties and operations” objectives hit, only Objective 1.3.f had more than 4% of total ACCUPLACER hits. The NAEP hits to “Number properties and operations” objectives were again more evenly distributed.

Despite the many differences, there were some commonalities across the two tests at the objective level. Both had at least 1% of hits to 13 objectives:

- 1.1.d, “Represent, interpret, or compare expressions for real numbers, including expressions using exponents and logarithms”
- 1.3.b, “Perform arithmetic operations with real numbers, including common irrational numbers”
- 1.3.f, “Solve application problems involving numbers, including rational and common irrationals”
- 1.4.d, “Solve multistep problems involving percentages, including compound percentages”
- 2.3.c, “Use the definitions of sine, cosine, and tangent as ratios of sides in a right triangle to solve problems about length of sides and measure of angles”
- 4.1.a, “Read or interpret graphical or tabular representations of data”
- 5.1.e, “Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or *trigonometric functions from tables, graphs, or equations”
- 5.2.a, “Create and translate between different representations of algebraic expressions, equations, and inequalities (e.g., linear, quadratic, exponential, or *trigonometric) using symbols, graphs, tables, diagrams, or written descriptions”
- 5.3.c, “Perform basic operations, using appropriate tools, on algebraic expressions including polynomial and rational expressions”
- 5.3.d, “Write equivalent forms of algebraic expressions, equations, or inequalities to represent and explain mathematical relationships”
- 5.3.e, “Evaluate algebraic expressions including polynomials and rational expressions”
- 5.3.f, “Use function notation to evaluate a function at a specified point in its domain and combine functions by addition, subtraction, multiplication, division, and composition”
- 5.4.a, “Solve linear, rational, or quadratic equations or inequalities, including those involving absolute value”

Clearly, the greatest degree of overlap was found in hits on both tests to objectives for “Algebra.”

Given the total numbers of objectives hit on each test (73 for NAEP and 56 for ACCUPLACER), the degree of overlap at the objective level is limited.

Overlap in content alignment to the ACCUPLACER framework can also be examined at the more finely grained objective level. Table 34 shows the overlap of alignment of each assessment to the ACCUPLACER framework in terms of the percentages of total hits.

Table 34. Summary of the Overlap of Content Alignment between NAEP and ACCUPLACER Items and the ACCUPLACER Framework at the Objective Level

			NAEP Items (164 items)		**ACCUPLACER Items** (45 items)	
			Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
Standards	**Goals**	**Objectives**	**% of Total Hits**		**% of Total Hits**	
A		A	0	0	0	0
	A.1	A.1.a	2	0	1	1
		A.1.b	0	0	2	2
		A.1.c	0	0	2	2
		A.1.d	0	0	1	0
		A.1.e	0	0	0	0
		A.1.f	0	0	0	0
		A.1.g	0	0	1	1
		A.1.h	0	0	2	2
		A.1.i	0	0	1	0
	A.2	A.2	0	0	0	0
		A.2.a	0	0	0	0
		A.2.b	1	1	1	1
		A.2.c	0	1	2	2
		A.2.d	0	0	3	2
		A.2.e	0	1	0	0
		A.2.f	0	0	2	2
		A.2.g	0	0	0	0
		A.2.h	0	1	2	4
		A.2.i	0	0	0	0
		A.2.j	0	0	2	2
		A.2.k	0	0	0	0
		A.2.l	0	0	0	0
	A.3	A.3.a	6	7	2	2
		A.3.b	1	2	2	2
		A.3.c	0	1	3	4
		A.3.d	6	7	4	5
		A.3.e	1	2	0	0
		A.3.f	11	9	0	0
		A.3.g	4	2	0	1

Standards	Goals	Objectives	NAEP Items (164 items)		ACCUPLACER Items (45 items)	
			Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
			% of Total Hits		**% of Total Hits**	
B	B.1	B.1.a	1	1	1	1
		B.1.b	0	0	2	2
		B.1.c	1	0	3	2
	B.2	B.2.a	5	4	4	4
		B.2.b	0	0	0	0
		B.2.c	1	1	4	4
		B.2.d	3	3	3	2
		B.2.e	0	0	0	0
		B.2.f	0	0	8	7
		B.2.g	0	0	0	0
		B.2.h	0	0	0	0
		B.2.i	1	1	3	2
		B.2.j	0	0	2	1
		B.2	0	1	0	0
	B.3	B.3.a	2	3	4	5
		B.3.b	1	1	1	2
		B.3.c	0	0	0	0
		B.3.d	4	4	0	0
		B.3.e	7	3	1	3
		B.3.f	6	4	2	0
		B.3	1	0	0	0
C		C	0	0	0	0
	C.1	C.1.a	0	0	0	0
		C.1.b	0	0	0	0
		C.1.c	0	0	0	0
		C.1.d	0	0	1	3
		C.1.e	0	1	0	0
		C.1.f	1	1	1	1
		C.1.g	0	0	0	1
		C.1	0	0	0	0
	C.2	C.2.a	0	0	1	2
		C.2.b	1	1	2	4
		C.2.c	0	0	1	1

Standards	Goals	Objectives	NAEP Items (164 items)		ACCUPLACER Items (45 items)	
			Panel 1	Panel 2	Panel 1	Panel 2
			% of Total Hits		% of Total Hits	
		C.2.d	1	2	0	0
		C.2.e	0	0	0	0
		C.2	0	0	0	0
	C.3	C.3.a	3	2	1	2
		C.3.b	1	3	2	1
		C.3.c	2	2	0	0
		C.3.d	0	0	0	0
		C.3.e	1	2	1	2
		C.3	1	0	0	0
	C.4	C.4.a	0	0	0	0
		C.4.b	0	0	0	0
		C.4.c	2	3	0	0
		C.4.d	0	0	0	0
		C.4.e	2	1	2	1
		C.4.f	0	0	0	0
		C.4.g	0	0	0	0
		C.4	0	1	0	0
	C.5	C.5	1	1	0	0
		C.5.a	0	0	0	0
		C.5.b	1	1	0	0
		C.5.c	3	6	1	0
		C.5.d	2	1	1	0
		C.5.e	1	1	0	0
		C.5.f	1	1	2	2
		C.5.g	1	1	2	2
		C.5.h	0	0	2	1
	C.6	C.6.a	0	0	2	2
		C.6.b	3	3	2	2
		C.6.c	0	0	0	0
		C.6.d	1	0	2	2
		C.6.e	1	1	0	0
		C.6.f	1	1	0	0
		C.6.g	0	0	0	0

Standards	Goals	Objectives	NAEP Items (164 items)		ACCUPLACER Items (45 items)	
			Panel 1	Panel 2	Panel 1	Panel 2
			% of Total Hits		% of Total Hits	
		C.6.h	0	0	0	0
		C.6	1	1	0	0

Percentages in table may not sum to 100% due to rounding.

In relation to the ACCUPLACER framework, the distribution of item alignments to objectives was relatively even on both tests, with most objectives receiving 1% and 4% of total hits from both sets of items. Only a few objectives received more than 4% of total hits on either or both tests. For alignment of NAEP to ACCUPLACER A, "Arithmetic," Objectives A.3.a ("Rate problems including ratio and proportion"), A.3.d ("Measurement problems"), and A.3.f ("Problems related to graphs and tables") each received over 4% of NAEP total hits from both panels. For ACCUPLACER B, "Elementary algebra," three objectives—B.2.a ("Evaluating formulas and other algebraic expressions"), B.3.e ("Solving verbal problems in an algebraic context including geometric reasoning"), and B.3.f ("Graphing")—each received over 4% of NAEP total hits from at least *one* panel. For ACCUPLACER C, "College level math," only Objective C.5.c ("Graphic properties, exponential and logarithmic functions") received over 4% of NAEP total hits from *one* panel.

For alignment of ACCUPLACER items to "Arithmetic," Objective A.3.d ("Measurement problems") received over 4% of ACCUPLACER total hits from one panel; for alignment to "Elementary Algebra," Objective B.2.f ("Factoring difference of squares") received over 4% ACCUPLACER total hits for both panels. No objective for "College level math" received more than 4% of ACCUPLACER total hits. In contrast to the results for the NAEP short-version items' alignment to the NAEP framework, there were more clusters of hits over 4% of total hits to specific ACCUPLACER objectives, particularly in "Arithmetic." However, the percentages of total hits for each objective were fairly small on both tests, with only one objective receiving over 7% of total hits.

Each assessment was found to cover objectives that the other did not. NAEP items had hits from both panels to ten objectives for which ACCUPLACER items had no hits from either panel.

ACCUPLACER objectives with hits from NAEP items only:

- A.3.e, "Average"
- A.3.f, "Problems related to graphs and tables"
- B.3.d, "Translating written phrases or sentences into algebraic expressions or equations"
- C.2.d, "Exponential equations"
- C.3.c, "Conics"
- C.4.c, "Series and Sequences"
- C.5.b, "Exponents and logarithms"
- C.5.e, "Composition of functions"

- C.6.e, “Trigonometric equations and inequalities”
- C.6.f, “Trigonometric identities”

The ACCUPLACER items had hits from both panels to 15 ACCUPLACER objectives to which the NAEP items had none from either panel:

- A.1.b, “Addition and subtraction of fractions or mixed numbers”
- A.1.c, “Multiplication involving fractions or mixed numbers”
- A.1.g, “Squares and square roots (whole numbers, fractions)”
- A.1.h, “Recognition of equivalent fractions and mixed numbers”
- A.2.d, “Division of a decimal by a whole number other than a power of ten or by a decimal”
- A.2.f, “Division of a decimal by a power of ten”
- A.2.j, “Percent one number is of another”
- B.1.b, “Operations with signed numbers”
- B.2.f, “Factoring difference of squares”
- B.2.j, “Division of monomials and polynomials including simplification of algebraic fractions”
- C.1.d, “Multiplication, division, and simplification of algebraic fractions”
- C.2.a, “Linear equations and inequalities”
- C.2.c, “Systems of equations and inequalities”
- C.5.h, “Periodicity, amplitude, and other properties”
- C.6.a, “Fundamental definitions of trig functions (e.g., sin, cos, tan, etc.)”

As indicated above and shown in Table 34, within “Arithmetic,” NAEP items had more coverage overall of objectives for Goal A.3 (“Applications”) than did the ACCUPLACER short-version items. The ACCUPLACER short-version items had more coverage of objectives for Goals A.1 (“Whole numbers and fractions”) and A.2 (“Decimals and percents”) than did the full NAEP item pool. For “Arithmetic,” therefore, ACCUPLACER items had a greater emphasis on objectives for Goals A.1 and A.2 while NAEP items had a greater emphasis on objectives for Goal A.3.

Within “Elementary algebra,” the results also varied across goals. Overall, the ACCUPLACER items had stronger, though still limited, coverage of Goal B.1 (“Integers and rationals”) and its objectives. ACCUPLACER items had hits to three of three objectives for this goal, while NAEP had less coverage, with hits to two of three Goal B.1 objectives.

The ACCUPLACER items had about twice as many hits as the NAEP items to Goal B.2 (“Algebraic expressions”) objectives. For this goal, NAEP had hits distributed among four of ten objectives, with Objectives B.2.a (“Evaluating formulas and other algebraic expressions”) and B.2.d (“Positive rational roots and exponents”) receiving the highest percentages of NAEP hits. ACCUPLACER items had hits distributed among six of ten Goal B.2 objectives, with Objectives B.2.a, B.2.c (“Multiplication of monomials and polynomials”), and B.2.f (“Factoring difference of squares”) receiving the most hits. Although Objectives B.2.a and B.2.d received similar emphasis in both tests, the ACCUPLACER items had a greater percentage of hits for more of the Goal B.2 objectives overall, with particular emphasis on Objective B.2.f.

NAEP items had approximately twice as many hits to Goal B.3 (“Equations, inequalities, and word problems”) objectives than did ACCUPLACER items. For objectives in Goal B.3, NAEP had fuller coverage, with hits distributed to five of six objectives and the greatest emphasis on Objectives B.3.d (“Translating written phrases or sentences into algebraic expressions or equations”), B.3.e (“Solving verbal problems in an algebraic context including geometric reasoning”), and B.3.f (“Graphing”). The ACCUPLACER items had hits to four of five objectives, with Objective B.3.a (“Solving linear equations and inequalities”) receiving the most emphasis.

Within “College level math,” both tests had very limited coverage of objectives for Goal C.1 (“Algebraic operations”). Both had hits to Objective C.1.f (“Powers, roots, radicals”); ACCUPLACER also had hits to Objective C.1.d (“Multiplication, division, and simplification of algebraic fractions”). For Goal C.2 (“Solution of equations and inequalities”), coverage was also limited on both tests, especially for NAEP, with only two of five objectives hit. Both NAEP and ACCUPLACER had hits to Objective C.2.b (“Quadratic equations”). ACCUPLACER items also had hits to Objective C.2.a (“Linear equations and inequalities”) and C.2.c (“Systems of equations and inequalities”), and NAEP had hits to Objective C.2.d (“Exponential equations”). The NAEP items had stronger coverage of Goal C.3 (“Coordinate geometry”) objectives. Both tests did, however, cover the same three objectives: C.3.a (“The coordinate plane”), C.3.b (“Straight lines”), and C.3.e (“Graphs of algebraic functions”).

ACCUPLACER short-version items had hits to only one (C.4.e) of seven objectives for Goal C.4 (“Applications and other algebra topics”). The NAEP items had more, though still limited, coverage with hits to both Objectives C.4.c (“Series and Sequences”) and C.4.e (“Permutations and combinations”). NAEP also had more coverage for Goal C.5 (“Functions”), with hits to six of eight objectives: C.5.b (“Exponents and logarithms”), C.5.c (“Graphical properties, exponential and logarithmic functions”), C.5.d (“Domain and range”), C.5.e (“Composition of functions”), C.5.f (“Inverse functions”), and C.5.g (“Computations with simple functions”). The ACCUPLACER items had 1% or more total hits to Objective C.5.f and C.5.g and to Objective C.5.h (“Periodicity, amplitude, and other properties”) C.5.c and C.5.d each received 1% of total ACCUPLACER hits from one panel only.) Finally, for Goal C.6 (“Trigonometry”), both NAEP and ACCUPLACER items had hits to the same one of eight objectives, with 3% of NAEP and 2% of ACCUPLACER hits to Objective C.6.b (“Right triangle trigonometry and circular functions”). The remaining NAEP hits were distributed to two objectives for which ACCUPLACER had no hits—C.6.e (“Trigonometric equations and inequalities”) and C.6.f (“Trigonometric identities”)—and to Goal C.6. The remaining ACCUPLACER hits were to objectives not hit by any NAEP items: C.6.a (“Fundamental definitions of trig functions e.g., sin, cos, tan, etc.”) and C.6.d (“Graphs of trigonometric functions”).

Overall, NAEP had somewhat stronger coverage of objectives within Goals C.3, C.4, and C.5; ACCUPLACER had slightly stronger coverage of objectives for Goal C.2. The two tests had similar coverage of Goals C.1 and C.6, though the distribution to specific objectives for these goals varied across tests.

Despite the many differences, there were some commonalities across the two tests at the objective level. Across all standards, the two tests had hits from both panels to 22 of the same objectives, with similar percentages (the same or within 1%) of hits to 13 of those objectives:

- A.2.b, “Multiplication of decimals by whole numbers other than powers of ten and by decimals”
- A.3.b, “Percent problems”
- B.1.a, “Ordering”
- B.2.a, “Evaluating formulas and other algebraic expressions”
- B.2.d, “Positive rational roots and exponents”
- B.3.b, “Systems of linear equations”
- C.1.f, “Powers, roots, radicals”
- C.3.b, “Straight lines”
- C.3.e, “Graphs of algebraic functions”
- C.4.e, “Permutations and combinations”
- C.5.f, “Inverse functions”
- C.5.g, “Computations with simple functions”
- C.6.b, “Right triangle trigonometry and circular functions”

Disregarding percentages, both tests also had hits from both panels to nine objectives:

- A.3.a, “Rate problems including ratio and proportion”
- A.3.d, “Measurement problems”
- B.2.c, “Multiplication of monomials and polynomials”
- B.2.i, “Operations with algebraic fractions involving addition, subtraction, multiplication, and division”
- B.3.a, “Solving linear equations and inequalities”
- B.3.e, “Solving verbal problems in an algebraic context including geometric reasoning”
- B.3.f, “Graphing”
- C.2.b, “Quadratic equations”
- C.3.a, “The coordinate plane”

As seen above and in Table 34, the greatest degree of overlap at the objective level was found in hits on both tests to nine of the same objectives for “Elementary algebra” and nine of the same objectives for “College level math.”

Although the differences between the two tests appear less pronounced in relation to the ACCUPLACER framework than in relation to the NAEP framework, this may be partly due to the more general language of the ACCUPLACER objectives compared to the more specific NAEP objectives. For example, the ACCUPLACER objective A.3.f (“Problems related to graphs and tables”) received a high percentage of total hits. The items coded to that objective would likely have been coded to several objectives within “Data analysis, statistics, and probability.” In one such case, the three NAEP items aligned to ACCUPLACER A.3.f that were also included in the sample for Sub-Study 1 were aligned to three different NAEP objectives (4.1.a, 4.1.d, and 4.2.d).

Categorical Concurrence

For alignment to the NAEP framework, the NAEP items used in the short version (42 items) were found to meet the typical WAT threshold value of at least six items for categorical concurrence for "Number properties and operations," "Measurement," "Geometry," and "Algebra." Categorical concurrence was not met for "Data analysis, statistics, and probability," although the standard received 4.5 and 5 mean hits. The ACCUPLACER items (105 items) met categorical concurrence for "Number properties and operations" and "Algebra," but not for "Measurement," "Geometry," or "Data analysis, statistics, and probability."

For alignment to the ACCUPLACER framework, the ACCUPLACER items used in the short version (45 items) were found to meet categorical concurrence for all standards. The NAEP items (164 items) also met categorical concurrence for all ACCUPLACER standards.

In reviewing whether the categorical concurrence threshold is met, it is important to consider the impact on this criterion of the number of items in the analyzed set (i.e., the more items that are analyzed, the more likely it is that the criterion will be met).

Depth-of-Knowledge Consistency and Range of Depth of Knowledge

For alignment to the NAEP framework, the NAEP items were found to meet depth-of-knowledge consistency in all standards. That is, for each standard, at least 50% of the items aligned to an objective in that standard were at or above the DOK level assigned to that objective. The ACCUPLACER items also met depth-of-knowledge consistency for all NAEP standards.

For alignment to the ACCUPLACER framework, DOK was analyzed as range of depth of knowledge. NAEP items that aligned to the ACCUPLACER framework were coded at DOK Levels 1–3, with most of the items at DOK Level 2. The percentages of NAEP items aligned to each ACCUPLACER standard by each panel follow:

- "Arithmetic," 25% and 19% at Level 1, 70% and 75% at Level 2, and 5% and 7% at Level 3
- "Elementary algebra," 37% and 42% at Level 1, 50% and 56% at Level 2, and 8% and 7% at Level 3
- "College level math," 31% and 26% at Level 1, 63% and 64% at Level 2, and 6% and 11% at Level 3

Almost all ACCUPLACER items aligned with the ACCUPLACER framework were coded at DOK Level 1 or Level 2. The percentages of Level 1 and Level 2 items aligned to each ACCUPLACER standard by each panel follow:

- "Arithmetic," 50% and 52% at Level 1 and 50% and 48% at Level 2
- "Elementary algebra," 71% and 74% at Level 1 and 29% and 26% at Level 2
- "College level math," 33% and 17% at Level 1 and 67% and 83% at Level 2

Range-of-Knowledge Correspondence

For alignment to the NAEP framework, the NAEP short-version set of items did not meet the criteria for range of knowledge for any standard (50% or more of objectives hit). No NAEP standard had more than 39% of its objectives hit by items in the short-version; "Geometry" had the most restricted range of knowledge, with only 14% of its objectives met. This result likely reflects the large number of objectives (130) relative to the number of items in the short form (42) used in this study. For alignment of the ACCUPLACER items (105) to the NAEP framework, range of knowledge was met for "Number properties and operations" and "Algebra," but was not met for "Measurement," "Geometry," or "Data analysis, statistics, and probability."

For alignment to the ACCUPLACER framework, the NAEP items (164 items) had a range of knowledge for "Elementary algebra" and "College level math." For "Arithmetic," range of knowledge was weakly met or not met, with 39% and 42% of its objectives hit. For the ACCUPLACER short-version set of items (45 items), range of knowledge was met for "Elementary algebra" (59% and 55% of objectives hit) and was weakly met for "Arithmetic" (47% of objectives hit). Range of knowledge was not met for "College level math," with 29% and 30% of objectives hit.

Balance of Representation

NAEP items in the short-version set had a balance of representation for all standards in the NAEP framework. The ACCUPLACER items had a balance of representation for "Measurement" and "Data analysis, statistics, and probability," but not for "Number properties and operations." The two panels differed on the results for "Geometry" and "Measurement." Panel 2 found that the ACCUPLACER items did not have balance of representation for these two standards, while Panel 1 found that the criterion for balance was met for "Geometry" and was weakly met for "Algebra."

In relation to the ACCUPLACER framework, the NAEP items met the criteria for balance of representation for "College level math" and weakly met the criteria for "Elementary algebra." They did not meet balance of representation for "Arithmetic." The ACCUPLACER short-version items met the criteria for balance of representation for all three ACCUPLACER standards.

Overall Conclusions

The following conclusions regarding the alignment of the 2009 NAEP Grade 12 Mathematics assessment and the ACCUPLACER Mathematics Core Tests can be drawn from the results of this alignment study.

What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by ACCUPLACER?

The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra (included in NAEP's "Number properties and operations" and "Algebra" standards and in ACCUPLACER's "Arithmetic," "Elementary algebra," and "College level math" standards), although their respective degrees of alignment and focus in these subdomains vary. Whereas the

NAEP items focus primarily on number or number operations and algebra content at the grade 12 level, with an emphasis on problem solving and application of concepts at that grade level, the ACCUPLACER items span a wider developmental and grade-level range (from basic to more advanced).

This difference in focus is consistent with the purposes of the two assessments and their frameworks. The NAEP objectives are written to describe assessable content for grade 12 mathematics; thus, the 130 objectives tend to address the skills and concepts specific to that grade. The purpose of ACCUPLACER is to help determine appropriate placement for an individual student, and so the 87 ACCUPLACER objectives are spread more broadly across grade levels and are intended to be more general.

To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on ACCUPLACER?

Regarding alignment to the NAEP framework, within the "Number properties and operations" and "Algebra" standards, NAEP items had broader overall coverage of the NAEP objectives than did ACCUPLACER. The 42 NAEP items (the short version used for within-framework alignment) aligned to 72 NAEP objectives, whereas the 105 ACCUPLACER items (one complete form of each of the three ACCUPLACER Mathematics Core Tests) aligned to only 56 NAEP objectives, with 44% of the ACCUPLACER item alignments aligning to only three NAEP objectives (all in "Number properties and operations" and "Algebra"). These differences in breadth and emphasis between the two assessments were evident across all NAEP standards. For example, in each assessment, items were aligned to four NAEP "Algebra" objectives for which the other assessment had no alignments, reflecting differences in emphasis within that standard.

Regarding alignment to the ACCUPLACER framework, ACCUPLACER items in the short version of 45 items covered all three standards—"Arithmetic," "Elementary algebra," and "College level math"—with a relatively even distribution, although "College level math" had the lowest percentage of item alignments. NAEP items in the full pool of 164 items also covered "Arithmetic," "Elementary algebra," and "College level math," with a fairly even distribution of approximately one-third of NAEP codable items aligned to each standard, although "Elementary algebra" received somewhat fewer item alignments. Despite these differences in emphasis, however, considering only codable items, the percentages of alignments to each ACCUPLACER standard were relatively evenly distributed in both assessments and similar in distribution across assessments. At the objective level, the distribution of item alignments to objectives was relatively even on both tests, although each assessment was aligned to some objectives to which the other was not.

In summarizing cross-framework alignment, there was somewhat less even distribution of items than observed in within-framework alignment. The majority of items on each test were found to align to objectives on the other test. However, the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from "Number properties and operations" in the NAEP framework, and four of the five goals in "Algebra." Conversely, the NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives.

Are there systematic differences in content and complexity between NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives (20 were rated uncodable by the majority of panelists in each panel). For other NAEP items that were aligned to an ACCUPLACER objective, there were often parts of those items not addressed by the objective. These items were coded as aligned, since they do assess an ACCUPLACER objective, but parts of the items also cover other skills not included in the ACCUPLACER framework.

Regarding differences in alignment of complexity, the items from both tests that aligned to the NAEP standards met the typical depth-of-knowledge (DOK) consistency threshold; that is, the items assessed the objectives at or above the DOK level of the objective. The items from both tests that aligned to the ACCUPLACER standards had somewhat different ranges of DOK. The ACCUPLACER short-version items were divided fairly evenly between Level 1 and Level 2. The NAEP items aligned to the ACCUPLACER framework had a wider range of DOK, with items at Level 1, 2, and 3, and a greater emphasis on Level 2 than was in the ACCUPLACER items.

VI. Discussion and Recommendations on Study Design

This alignment study involved the implementation of a study design custom-developed by Dr. Webb. Given the relatively early stage of the field of assessment-to-assessment alignment, and at the request of the Governing Board, this section includes some considerations and recommendations related to implementation of the study design during the pilot study and the operational studies (NAEP–ACCUPLACER reading and mathematics, and NAEP–SAT reading and mathematics). Process recommendations from the pilot study are included in Section II of this report and in the Pilot Study Report. In addition, some of the recommendations from the Pilot Study Report are restated here, as they relate to the overall study design. Except where specifically related to mathematics or ACCUPLACER, or otherwise stated, considerations and recommendations in this section are applicable to all four alignment studies.

Framework Selection

The selection of the framework document for use in an alignment study is a critical decision impacting the study logistics, results, and interpretation of findings. In short, the focus of a study is defined by the content of the framework used. In order to create the most complete description of the alignment of the two assessments, it is important, as was done in this NAEP–ACCUPLACER study, to acquire the most complete, detailed framework available, and then to select the most appropriate grain size for coding and analysis.

In this study, WestEd received from the Governing Board and the College Board very different framework documents with different levels of specificity of content for NAEP and ACCUPLACER. Among the most substantive differences was the NAEP framework's inclusion of language describing how students would apply the knowledge and skills, while the ACCUPLACER framework focused on content topics.

In interpreting the study results, it is also important to consider that panelists were selecting from 130 NAEP objectives and 87 ACCUPLACER objectives. Each framework had some internal overlap among the content of the objectives; indeed, in the domain of mathematics there is overlap among content, skills, and approaches necessary to solve problems. The large number of objectives in each framework affected a number of areas of the study. While it provided a detailed grain size for specifying the content, it increased coding time, since there were more objectives to analyze. It also increased the likelihood that discrepancies in coding items to objectives would need to be discussed. Finally, it increased the likelihood that some objectives would not be matched to one or more items.

Background Information on the Assessments

As described earlier, prior to the study, panelists received a required reading packet of information about the two assessments, including the full 2009 NAEP framework and background information about the ACCUPLACER assessment. During the study, additional review and discussion of aspects of the content of the full framework were provided for panelists to help them understand the coding documents in their complete context. For example, the full NAEP framework contextualizes some terms that appear in the standards and objectives used for alignment coding. For future studies, it may be beneficial to determine across studies what information panelists will learn sufficiently through advance reading, and what warrants

clarification or reinforcement during in-person training and discussion. This could inform further refinements to pre-study communication with panelists and the panelist training.

Depth of Knowledge Levels

Per the design document, Webb's depth of knowledge levels were applied as the criteria for cognitive complexity. In practice, panelists requested some clarification related to the interpretation and application of the criteria to grade 12 mathematics. In particular, there was some discussion among panelists and facilitators about the difference between problems requiring multiple computation steps (that is, the application of the same or similar concept multiple times), and those problems involving multiples concepts, and whether the former should be considered as DOK Level 1 or Level 2 for grade 12 students. In other cases, the clarity of the wording of the DOK level descriptions prompted discussion about appropriate interpretation.

In this study, the full range of DOK levels was not found in the items or objectives for either assessment. In Webb's DOK level descriptors, Level 4 is defined by the key elements of higher-order thinking and extended time. Under this definition, DOK Level 4 is only assigned to standards or tasks that describe knowledge and skills embodying higher-order thinking and that can only be demonstrated over time. This is not typically an expectation for a reading or mathematics assessment, even with the extended constructed-response item types found on NAEP. The importance of having both factors (higher-order thinking and extended time) in order to code an objective or item as Level 4 was included in the training and reflected in the discussions with facilitators. As a result, panelists found that they were not able to use DOK Level 4, effectively reducing the DOK choices to Levels 1–3.

Issues such as these suggest that examining the utility of the DOK levels for 12th grade preparedness may be useful. Such an examination would consider whether this configuration is warranted for use in future preparedness studies, or whether revision or extension would be advisable. If it is found that the DOK levels are most applicable to 12th grade preparedness in their current form, the Governing Board may wish to consider whether the assessments should be expanded in the future to include the capacity to measure knowledge and skills across the full four-level range of DOK.

Order of Sub-Studies

As described in Section II of this report, WestEd recommended and, receiving the Governing Board's approval, implemented a change of sub-study order, so that within-framework activities for each assessment would be completed prior to conducting the cross-framework analysis. The purpose of this change was to ensure that panelists would align an assessment's items to its own framework before being exposed to that framework through cross-framework item alignment. Coding the Pexam assessment items prior to the Pexam framework, as in the original order, could have risked limiting panelists' interpretation of the possible DOK of that framework's objectives to the objectives' operationalization in the item pool provided. In practice, this refinement to the design was effective and is recommended for future assessment-to-assessment alignment studies.

Placement of Correct Answers in Item Booklets

The item booklets reviewed by the panelists included each item's correct answer on that item's page. Panelists were instructed to answer the questions or solve the problems as a student would, but for some panelists the correct answer was a minor distraction that might have influenced their coding, and during the final debrief discussion, some panelists expressed that they would have preferred to have the correct answer hidden or provided on a separate page. Conversely, other panelists reported that the correct answer was useful and efficient in its location, and that they had no concerns about distraction. Given the potential distraction, and in an effort to present the items as closely as possible to the way students would experience them, the correct answers should be available separately from the items in future studies. Including this specification in the study design will help to ensure a standardized format across studies.

Cross-Panel Adjudication

The study design outlined the parameters for adjudication by replicate panels according to the four criteria. In practice, WestEd's development of an adjudication workbook facilitated this process greatly, providing all relevant data from each panel in a single sheet, with discrepant ratings flagged for facilitator review. Given the aggressive timeline for the studies, this increase in efficiency was important, and such a tool is recommended for future studies of this nature and scope.

Initial readings of the design document suggested that the outcome of the cross-panel adjudication process was to bring the two panels closer in the areas for which they were discrepant. Because of the interrelated nature of the alignment criteria (e.g., a discrepancy in depth-of-knowledge consistency can be the product of multiple factors, including match to objective and depth of knowledge), identifying all related items and then working with both panels to address the issue was a significant challenge. An early conversation with the COR clarified that the goal of the adjudication process was understanding the differences between the panels' results, particularly whether they were systematic or random, and not requiring the resolution of all such differences. This was an important clarification in the purpose of the replicate panel structure and the data this structure would produce, and it should be clarified in the design document for future use.

Data Analysis

The study design clearly outlines the process for alignment of each assessment to each framework, and recommends the WAT for this purpose. However, the design does not specify how the four separate sub-studies should be analyzed to determine the cross-assessment alignment. Thus, WestEd requested guidance in how the bi-directional framework analysis should be synthesized for reporting across assessments. In order to determine the most effective and meaningful method for analyzing the assessment-to-assessment alignment, the Governing Board hosted conversations with Dr. Webb, WestEd, and ACT. A representative from the College Board also attended to represent that organization on questions of data security. The analysis and presentation format presented in this report is the outcome of those discussions.

Another issue related to data analysis that required follow-up discussion was how to use the replicate panel data. The design document indicates that the results could be aggregated or

averaged once it was established that the panels were indeed replicate. However, the WAT system is not currently programmed to combine studies in this way. Following discussions with the Governing Board, Dr. Webb, and ACT, it was decided to report both panels' results separately in order to show areas where the replicate panels produced discrepant results, which may in itself be an interesting finding regarding alignment.

Other Factors That May Affect Alignment

The alignment methodology used in this study captures the degrees of alignment between the assessments and their respective framework in terms of content and cognitive complexity. However, it is important to consider alignment outcomes in light of other factors in the assessments, as summarized in Table 1 and in the Interim Report, and as mentioned in several panelists' evaluation forms. Among these other factors are reading difficulty, item type, item difficulty, and test purpose. For example, in reading specifically, although items may be aligned to the same objectives, the amount and level of reading (not just genre) may be an important difference between the two assessments in how they assess reading and 12th grade preparedness. Similarly, it is possible that there are other preparedness-related differences between the content of the assessments—related, for instance, to the variety of item types on NAEP (i.e., multiple choice, short constructed response, extended constructed response) in comparison with the single type on ACCUPLACER (i.e., multiple choice)—that extend beyond those differences that would be apparent from the alignment to each framework. In short, it is important to consider these alignment data in the context of the entire study, including the qualitative comparative analysis. Finally, when making comparisons of content and depth, it is important to keep in mind each assessment's purpose and use.

Timing and Panelist Workload

Based on lessons learned from the pilot study and an expectation of aggressive timelines, the study team implemented a number of processes to maximize efficiency of use of panelists' time. WestEd developed its adjudication workbook to quickly provide the cross-panel comparison information required for adjudication. Also, the replicate panels analyzed a reduced item pool for the NAEP items and a reduced ACCUPLACER pool for the mathematics alignment workshop. As a result, all panelists from reading and mathematics completed all study activities, with the reading panelists completing the study work in less than the allotted time. However, timing was closely linked to quantity of items and objectives, and, as described in this report and WestEd's comprehensive report on the NAEP–SAT mathematics study, this presented a challenge to keeping the mathematics panels to the allotted schedule. Panelist feedback on the long coding sessions confirms that, for mathematics, either a reduction in workload or an extension of the length of the workshop would be desirable. Therefore, monitoring overall workload should be an explicit objective of the study design.

Panelist Experience

Based on panelist evaluation survey responses, as well as in-person and email feedback, most panelists found the experience of serving on an alignment panel to be a rewarding one. The facilitators' content knowledge and their ability to efficiently and effectively manage group adjudication discussions were mentioned numerous times as being central to this positive experience, as were the effective planning and implementation of the workshop logistics.

An additional outcome of the study, mentioned by a number of panelists, was the professional development of being engaged in the interesting work of item alignment with a strong and diverse team of fellow professionals. For many panelists, it is an uncommon occurrence to spend a week discussing content with a team that might include high school teachers, university professors, and national consultants. Although the work was cognitively demanding and time-intensive, the opportunity for the panelists to discuss and apply their area of content expertise to a project they felt was of national importance was appreciated. Additionally, panelists tended to bond throughout the week, often dining together in the evenings. While it was not the purpose of the study, it is important that panelists found the experience worthwhile and rewarding to the extent that they remained engaged through the course of the study. This was certainly the case, and several panelists have asked to be considered for future alignment opportunities.

VII. References

The College Board. (2007). *ACCUPLACER® sample questions for students. Revised December 2007.* Retrieved March 29, 2010, from http://www.collegeboard.com/prod_downloads/student/testing/accuplacer/accuplacer-sample.pdf

The College Board. (2009a). *ACCUPLACER®.* Retrieved December 22, 2009, from http://professionals.collegeboard.com/higher-ed/placement/accuplacer

The College Board. (2009b). *ACCUPLACER®. Revealing potential. Expanding opportunity.* New York, NY: Author.

The College Board. (2010). *ACCUPLACER®. Introduction for students.* Retrieved March 29, 2010, from http://www.collegeboard.com/student/testing/accuplacer/?print=true

National Assessment Governing Board. (2008). *Mathematics Framework for the 2009 National Assessment of Educational Progress.* Developed for the National Assessment Governing Board in support of Contract No. ED-00-CO-0115, U.S. Department of Education, by the Council of Chief State School Officers, with subcontracts to the Council of Basic Education and the Association of State Supervisors of Mathematics and Grade 12 preparedness objectives developed under contract with Achieve, Inc.

National Assessment Governing Board. (2009a). *Content alignment studies of the 2009 National Assessment of Educational Progress (NAEP) for grade 12 reading and mathematics with the SAT and ACCUPLACER assessments of these subjects (Solicitation No. ED-NAG-09-R-0005).* Washington, DC: Author.

National Assessment Governing Board. (2009b). *Design of content alignment studies in mathematics and reading for 12th grade NAEP and other assessments to be used in preparedness research studies.* Washington, DC: Author.

National Assessment Governing Board (2009c). *Making new links,12th grade and beyond: Technical panel on 12th grade preparedness research, final report*. Washington, DC: U.S. Government Printing Office.

National Center for Educational Statistics. (2005). *NAEP questions tool.* Retrieved March 29, 2010, from http://nces.ed.gov/nationsreportcard/itmrlsx/landing.aspx

National Center for Educational Statistics. (2009). *Sample questions, grade 12, 2009. Mathematics. Reading. Science.* Retrieved March 29, 2010, from http://nces.ed.gov/nationsreportcard/pdf/demo_booklet/09SQ-G12-MRS.pdf

U.S. News & World Report. (2010). *Best colleges 2011*. Retrieved January 25, 2010, from http://colleges.usnews.rankingsandreviews.com/best-colleges

Webb, Norman L. (2005). *Web Alignment Tool (WAT) training manual.* Wisconsin: Author.

WestEd. (2010a). *Comprehensive Report: Alignment of 2009 NAEP Grade 12 mathematics and SAT mathematics* (Unpublished report submitted to the National Assessment Governing Board, Contract no. ED-NAG-09-C-001).

WestEd. (2010b). *Comprehensive Report: Alignment of 2009 NAEP Grade 12 reading and ACCUPLACER reading comprehension* (Unpublished report submitted to the National Assessment Governing Board, Contract no. ED-NAG-09-C-001).

WestEd. (2010c). *Comprehensive Report: Alignment of 2009 NAEP Grade 12 reading and SAT critical reading* (Unpublished report submitted to the National Assessment Governing Board, Contract no. ED-NAG-09-C-001).