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.

Comprehensive Report: Alignment of 2009 NAEP Grade 12 Reading and ACCUPLACER Reading Comprehension

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

Appendices Part 1

Appendices Part 2: Confidential and Proprietary

List of Tables

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:
Karen Anderson
John Fortier

Study Panelists:
[redacted]

Technical Advisor:
Norman Webb

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

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 Reading and ACCUPLACER Reading Comprehension

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 entry/placement in job training courses. The alignment study discussed in this report—the first of four comprehensive reports to be submitted to the Governing Board—evaluated the alignment between the NAEP and ACCUPLACER assessments in reading.

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 reading framework and the ACCUPLACER reading specifications, conducted between late 2009 and 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 reading framework, 2) the ACCUPLACER assessment and the ACCUPLACER reading specifications, 3) the grade 12 NAEP and the ACCUPLACER reading specifications, and 4) the ACCUPLACER assessment and the NAEP reading 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. Alignment criteria used and reported on in this study included categorical concurrence, depth-of-knowledge consistency, range of knowledge, and balance of representation.

This report addresses the following specific questions:

- What is the correspondence between the reading content domain assessed by NAEP and that assessed by ACCUPLACER?
- To what extent is the emphasis of reading 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 reading 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 (summarized here at the level of each framework's standards).

NAEP Framework Standards

1. "Locate/Recall"
2. "Integrate/Interpret"
3. "Critique/Evaluate"

ACCUPLACER Framework Standards

1. "Identifying main ideas"
2. "Direct statements/secondary ideas"
3. "Inferences"
4. "Applications"
5. "Sentence relationships"

NAEP Assessment to NAEP Framework Alignment

The NAEP short-form items (40) were found to assess all three NAEP standards. Of these three standards, "Integrate/Interpret" received the majority of alignments. The remaining NAEP items were distributed to "Locate/Recall" and "Critique/Evaluate," with the latter receiving the lowest number of alignments.

ACCUPLACER Assessment to NAEP Framework Alignment

The 55 ACCUPLACER items were found to assess two of the three NAEP standards: "Locate/Recall" and "Integrate/Interpret." One panel found that a few ACCUPLACER items assessed the third NAEP standard, "Critique/Evaluate"; the other panel found that no ACCUPLACER items assessed this standard. "Integrate/Interpret" received the majority of ACCUPLACER item alignments while "Locate/Recall" received a smaller percentage of the alignments.

ACCUPLACER Assessment to ACCUPLACER Framework Alignment

The 55 ACCUPLACER items were found to assess all of the five ACCUPLACER standards. "Inferences" received the greatest number of alignments, closely followed by "Sentence Relationships" and "Direct statements/secondary ideas." "Identifying main ideas" and "Applications" each received somewhat fewer alignments.

NAEP Assessment to ACCUPLACER Framework Alignment

NAEP items from the complete pool (131 items) were found to assess four of the five ACCUPLACER standards. “Inferences” received the greatest number of alignments, followed by “Direct statements/secondary ideas” and “Applications.” “Identifying main ideas” received the fewest alignments. No NAEP items were found to assess “Sentence Relationships.” Additionally, 12% and 15% of the NAEP items were judged by the majority of panelists in each panel to not align to the ACCUPLACER framework.

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 NAEP items in the short version met categorical concurrence for “Locate/Recall” and “Integrate/Interpret.” Categorical concurrence was not met for “Critique/Evaluate,” although it approached the threshold. The ACCUPLACER items met categorical concurrence for the two standards to which they were aligned, “Locate/Recall” and “Integrate/Interpret,” but not for “Critique/Evaluate.”

For alignment to the ACCUPLACER framework, the ACCUPLACER items met categorical concurrence for all five ACCUPLACER standards. The NAEP items met categorical concurrence for the four standards to which they aligned but not for “Sentence Relationships,” to which no items were aligned.

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).

Depth-of-Knowledge Consistency

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 met depth-of-knowledge consistency for all standards. The ACCUPLACER items met depth-of-knowledge consistency only for “Locate/Recall.” Both panels found that the majority of ACCUPLACER items aligned to “Integrate/Interpret” had a lower DOK level than that of the standard. The ACCUPLACER items had minimal to no alignments to “Critique/Evaluate.”

For alignment to the ACCUPLACER framework, the ACCUPLACER items were found to meet depth-of-knowledge consistency in all objectives. The NAEP items met depth-of-knowledge consistency in the four objectives to which there were alignments.

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, for both NAEP and ACCUPLACER, only “Integrate/Interpret” had a range of knowledge, with 50% or greater of the 17 objectives within that standard receiving alignments.

For alignment to the ACCUPLACER framework, the ACCUPLACER items had hits to all five standards and NAEP items had hits to four of five standards. However, because the ACCUPLACER framework has only one level, the range of knowledge analyses are not applicable for this framework.

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 had a balance of representation for each of the two NAEP standards to which they were aligned.

The ACCUPLACER specifications have only one level, so the balance of representation analyses are not applicable for these specifications.

Overall Conclusions

The following conclusions regarding the alignment of the 2009 NAEP Grade 12 Reading and the ACCUPLACER Reading Comprehension test can be drawn from the results of this alignment study.

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

The greatest commonality between the two tests is in their shared emphasis on the broad skills of comprehending and interpreting informational text, primarily through inferential reasoning. This is evident in the majority of items on both tests (two-thirds to three-fourths) matched to the NAEP standard “Integrate/Interpret: Make complex inferences within and across texts.” On both tests, the majority of alignments to “Integrate/Interpret” were to objectives that apply to informational text only or across both informational and literary texts.

The shared emphasis on the comprehension and interpretation of informational text can also be seen in the alignments on both tests to the ACCUPLACER framework. Although the ACCUPLACER standards do not explicitly refer to text type, they focus almost exclusively on elements typical of informational text. A majority of both NAEP and ACCUPLACER items were matched to the ACCUPLACER standard “Inferences,” and both tests had notable percentages of alignments to “Direct statements and secondary ideas” and “Applications.” A smaller percentage of items on both tests were aligned to “Identifying main ideas.”

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

As previously discussed, the alignments both within and across frameworks show that both tests emphasize the comprehension and interpretation of informational text, particularly through the use of inference. Within this broad area of convergence, however, there are differences in

emphasis revealed in the alignments to specific objectives within both frameworks. In relation to the NAEP framework, the NAEP short-version items showed a far greater emphasis on the comprehension of vocabulary in context (Objective 4.a) and on the analysis of an author's use of language (Objective 1.d). In relation to the ACCUPLACER framework, NAEP items showed more emphasis on the use of inference to interpret text ("Inferences"). The higher percentage of NAEP items aligned to "Applications" also reflects the greater emphasis in NAEP on understanding authors' use of language.

In relation to the ACCUPLACER framework, the ACCUPLACER items showed a greater emphasis than the NAEP items on the identification of main ideas. In relation to the NAEP framework, the ACCUPLACER items showed more emphasis on the recall of specific details, facts, and information (NAEP 1.1.a).

In general, in the cross-framework alignments, the matches found in each test to the other's framework (NAEP to ACCUPLACER and ACCUPLACER to NAEP) tended to be for the most general objectives within that framework. For example, the great majority of hits for ACCUPLACER items to NAEP objectives for "Integrate/Interpret" were to two of the most broadly stated NAEP objectives, "Draw conclusions" (2.3.b) and "Compare or connect ideas" (2.1.b). Many of the more specific NAEP objectives for "Integrate/Interpret," such as "Find evidence in support of an argument" (2.2.c), received far fewer or no hits from ACCUPLACER items. Compared to ACCUPLACER, the NAEP items were more evenly distributed among NAEP objectives.

The majority of alignments for NAEP items to ACCUPLACER standards were also to the broadest of those standards—"Inferences" and "Applications," both of which overlap in content with a number of NAEP objectives but at a higher level of generality. The more specific ACCUPLACER standard, "Identifying main ideas," received far fewer alignments from NAEP items.

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 reading subdomains are missing or not aligned?

In regard to differences in content, NAEP addresses reading skills related to both literary and informational text, while ACCUPLACER does not address reading skills specific to literary text. As expected, based on the framework-to-specifications comparison in the Interim Report, ACCUPLACER items had minimal matches to NAEP objectives for literary text. The main area of alignment of ACCUPLACER items to the NAEP framework, NAEP objectives in "Locate/Recall" and "Integrate/Interpret," applied to informational text only or to both informational and literary text.

The ACCUPLACER items also had minimal to no coverage of the NAEP standard "Critique/Evaluate." These findings are also consistent with the comparison of the two frameworks in the Interim Report; overall, the language of the ACCUPLACER objectives ("understand," "comprehend," "recognize") places more emphasis on comprehension and interpretation of text ("distinguish the main idea from supporting ideas" or "perceive connections

between ideas made—implicitly—in the passage") than on critical analysis or evaluation ("Evaluate the strength and quality of evidence used by the author to support his or her position" in NAEP Objective 3.3.b, or "Judge the author's craft and technique" in NAEP Objective 3.1.a).

In regard to complexity, both assessments were found to meet the criteria for depth of knowledge consistency in relation to their own framework. In relation to the NAEP framework, however, only the NAEP items met the criteria for DOK consistency for all NAEP standards. The ACCUPLACER items met the criteria for depth of knowledge consistency only for NAEP "Locate/Recall." Although the majority of the ACCUPLACER item alignments were to objectives for NAEP "Integrate/Interpret," over half of these items were found to have a DOK level below that of the standard. In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the more in-depth reasoning required by DOK Level 3. NAEP, by contrast, includes much longer reading passages and both multiple-choice and constructed-response items.

NAEP covers skills specific to the comprehension and analysis of literary text while ACCUPLACER does not. In addition, NAEP covers the skills of evaluating and critiquing text, skills not addressed by ACCUPLACER. Finally, NAEP has a wider range of cognitive complexity than ACCUPLACER, with a substantially higher percentage of items at DOK Level 3, requiring more in-depth analysis or evaluation. However, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

Overall, the NAEP items covered a broader range of cognitive complexity than the ACCUPLACER items. This is also apparent in the frameworks. The three NAEP standards, defined in terms of three different "cognitive targets" ("Locate/Recall," "Integrate/Interpret," and "Critique/Evaluate"), cover a broader range of cognitive complexity supported by the use of longer reading passages and the inclusion of both short and extended constructed-response items. The language of the ACCUPLACER standards ("understand," "comprehend," "recognize") places more emphasis on comprehension and interpretation of text (e.g., "distinguish the main idea from supporting ideas" in ACCUPLACER A, "Identifying main ideas," or "perceive connections between ideas made—implicitly—in the passage" in ACCUPLACER C, "Inferences") than on critical analysis or evaluation (e.g., "Evaluate the strength and quality of evidence" in NAEP 3.3.b, or "Judge the author's craft" in NAEP 3.1.a). In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the cognitive complexity typical of DOK Level 3 items. Although the NAEP items show a greater range of cognitive complexity and a greater emphasis on critical thinking, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

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 the state of 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 in addition to 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 NAEP Grade 12 Reading (NAEP) and the ACCUPLACER Reading Comprehension test (ACCUPLACER). Alignment findings from the studies of the alignment between NAEP and ACCUPLACER Mathematics Core Tests, SAT Critical Reading, and SAT Mathematics are presented in separate reports (WestEd, 2010a, 2010b, 2010c).

Governing Board's Approach to Preparedness

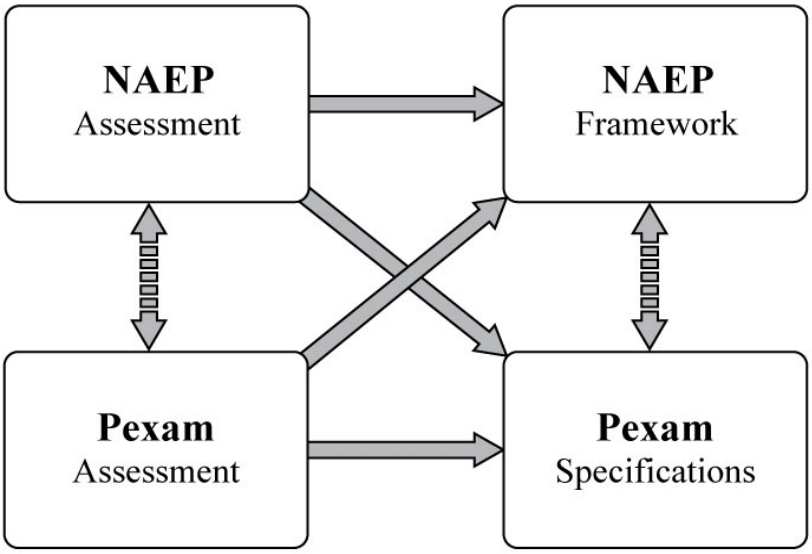
The Governing Board is focusing its conceptualization of 12th grade preparedness on academic qualifications and does not propose to address a range of behavioral and attitudinal aspects of student performance in postsecondary activities that are not measured by NAEP (e.g., time management skills, diligence). 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[1] (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. (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 reading framework and the ACCUPLACER reading 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[2]



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

[2] 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.

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 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 NAEP and ACCUPLACER measure the reading 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 reading 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 reading content domain assessed by NAEP and that assessed by ACCUPLACER?
- To what extent is the emphasis of reading 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 reading subdomains are missing or not aligned?

Alignment Study

The NAEP–ACCUPLACER reading 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. It comprised a series of four sub-studies to determine the degree of alignment between 1) the grade 12 NAEP and the NAEP reading framework, 2) the ACCUPLACER assessment and the ACCUPLACER reading

specifications, 3) the grade 12 NAEP and the ACCUPLACER reading specifications, and 4) the ACCUPLACER assessment and the NAEP reading 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 reading content experts, each comprising seven 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 reading alignment study was conducted concurrently with the NAEP–ACCUPLACER mathematics 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 reading alignment study for these two assessments (see Section III of this report for alignment results).

The development of the NAEP reading 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.[3] The development of the ACCUPLACER reading 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.

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 reading;
- Section III presents the results of this study;
- Section IV presents results of panelists' evaluation of the process;
- Section V presents a summary of results and conclusions;

[3] 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.

- Section VI presents a contractor 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 includes an overview of the components of the study design, followed by a detailed description of the methodology and study procedures. The methodology, procedures, and logistics described in this section reflect lessons learned from the pilot alignment study of the NAEP–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. This section also describes the specific elements of the methodology, procedures, and logistics that were modified as a result of the pilot study.

Study Design Overview

This subsection 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. 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 reading assessment aligns in content and complexity to the ACCUPLACER reading 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 reading 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 reading 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 framework (cross-framework). This within-framework baseline alignment was important in interpreting the degree of cross-framework alignment.

The alignment methodology employed in this study called for each objective to be assigned a DOK level, 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.

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

1. NAEP framework coded for DOK
2. NAEP items coded to NAEP framework
3. ACCUPLACER framework coded for DOK

4. ACCUPLACER items coded to ACCUPLACER framework
5. NAEP items coded to ACCUPLACER framework
6. ACCUPLACER items coded to NAEP framework

These item-level codes were then analyzed at the test level to produce reports of categorical concurrence, depth-of-knowledge consistency, range of knowledge, and balance of representation.

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 reading panels were required to reach joint agreement on the DOK levels of each assessment framework's objectives.[4] Following each panel's individual coding of objectives' DOK levels, the facilitators met to identify and discuss discrepancies. Prior to alignment coding of each assessment's items, each panel independently coded that assessment's 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.

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 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 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 changes to their codes if their judgment of the coding had changed. 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

[4] 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" (p. 13).

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's 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), 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.

This alignment workshop was conducted the week of April 12–16, 2010, at the Westin Grand hotel in Washington, DC, and engaged two replicate panels of reading content experts, each comprising seven 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 later in this section of the report.

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. Panelists assigned DOK levels to each of the framework's objectives, discussing as needed to reach agreement across panels on these ratings as described earlier; panelists then independently assigned DOK levels and determined objective alignments for each test item. The WAT was the primary data collection tool and was used subsequently for data analysis and report generation. Panelists were also encouraged to record all alignment ratings in their item books as a backup against technical issues.

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 C and are discussed in Section III of this report.

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. 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 from the pilot study follows.

Sequence of Study Steps

- Modify the coding order to code DOK levels of both 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 that 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 (examples of modifications are described later in this section).
- Identify and provide additional information, if available, to elaborate on the ACCUPLACER framework used for coding.[5]

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 in advance of the study, 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 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).
- 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.

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

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.

All recommendations were implemented.

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 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.

Dr. Karen Anderson served as lead 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 reading 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. Dr. Anderson has worked in K–12 and higher education, both public and private, for over 25 years, as a teacher, writer, assessment developer, and English language arts/reading content specialist. For the past four years, Dr. Anderson has worked with WestEd, specializing in the areas of English language arts/reading standards and assessment at national, state, and local levels. In particular, she has served as English language arts content lead and content analyst on numerous alignment studies, responsible for the overall quality of reading analyses, including the training of raters, facilitation of calibration discussions, and drafting of reports. Her alignment work has included both fixed-form and computer-adaptive assessments. Dr. Anderson received a BA in English, cum laude, from the California State University, Stanislaus, and an MA and PhD in English from the University of California, Davis.

Mr. John Fortier served as the second reading 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. Working with Dr. Norman Webb, Mr. Fortier has led approximately 45 English language arts alignment studies for 25 states, Puerto Rico, and the country of Qatar. Mr. Fortier taught English, speech, and debate in high schools and colleges before going to the Wisconsin Department of Public Instruction as a consultant in language arts assessment. In 1997, he was appointed Wisconsin's Assistant State Superintendent for Instructional Services, which included curriculum, assessment, and teacher education and licensing. While in that position, he served as staff to the Governor's Commission on Model Academic Standards and supervised the development of educational standards for the state of Wisconsin. He has also served as consultant for a number of states and testing companies. Mr. Fortier holds a BS and an MS in education from the University of Wisconsin, Madison.

Panel Criteria for Recruitment and Panelist Qualifications

A total of fourteen panelists, seven 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. One member of one panel reported general exposure to NAEP through involvement in prior item anchoring studies for a different NAEP reading assessment; this exposure was deemed by the Governing Board COR to not be problematic for the purposes of this alignment workshop.

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

- Referrals from the NAEP Reading Framework Planning Committee (2009), as identified in the 2009 framework.
- 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 English; the College English Association; the Two-Year College English Association; the International Reading Association, the Reading Teacher Editorial Council, and the Journal of Adolescent and Adult Literacy Editorial Council.
- Departments of English and schools of education from top-ranked colleges and universities across the country.[6]

Panels were balanced in numbers of representatives from secondary and postsecondary settings:

- On the first panel, 57% of panelists (4 of 7) reported experience in both secondary and postsecondary reading education; 29% (2 of 7) had secondary teaching experience only; and 14% (1 of 7) had postsecondary teaching experience only.
- On the second panel, 43% of panelists (3 of 7) reported experience in both secondary and postsecondary reading education; 29% (2 of 7) had secondary teaching experience only; and 29% (2 of 7) had postsecondary teaching experience only.

The composition of panels was balanced according to background expertise and experience with the NAEP and ACCUPLACER assessments (including both current and prior experience). 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 five women on each of the seven-member panels,

[6] 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.

as was representation of advanced degrees, with six doctoral degrees represented on each panel. Panelists represented a range of geographic areas, including the Northeast (New York), the Southeast (Florida, Georgia), the Midwest (Illinois, Kansas, Minnesota, Missouri), the Southwest (Texas), the Northwest (Montana, Washington, Wyoming), and the West (Arizona, California). WestEd was unable to achieve race/ethnicity diversity, however. All panelists but one on each panel identified themselves as White/Caucasian/of European descent; the other two panelists identified themselves as Asian/Caucasian (1) and Native American or Alaskan (1). A list of panelists organized by panel follows.

Reading Panel 1

[redacted]

Reading Panel 2

[redacted]

Standards and Representation of the Reading 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), a consultant to the Governing Board (Dr. Karen Wixson), and ACT to ensure that a NAEP reading framework organization appropriate for use in alignment studies was implemented. The form of the NAEP reading framework approved for this operational study was based on a version of Exhibit 8 ("Cognitive targets") of the Governing Board's *Reading Framework for the 2009 National Assessment of Educational Progress* (National Assessment Governing Board, 2008, p. 39) that ACT used in the alignment study for NAEP and WorkKeys. For that study, ACT adapted Exhibit 8 into a standard/goal/objective organizational structure, using Exhibit 8 columns ("Locate/Recall," "Integrate/Interpret," and "Critique/Evaluate") as standards and Exhibit 8 rows ("Both Literary and Informational Text," "Specific to Literary Text," and "Specific to Informational Text"), as goals, converting the content of each Exhibit 8 cell into discrete objectives. This organizational structure was provided to WestEd by the Governing Board for use in WestEd's NAEP–ACCUPLACER study.

In addition, based upon feedback from WestEd's NAEP–ACCUPLACER reading pilot study and through the course of discussions between WestEd, the Governing Board, and ACT, additional refinements were made to the content of ACT's NAEP framework organizational structure. The objective of these refinements was to capture the intent of the *Reading Framework for the 2009 National Assessment of Educational Progress* as fully as possible while reducing ambiguities and redundancies and maximizing consistency across the standards and objectives. One focus of discussion was the choice of verbs to use in constructing standard, goal, and objective statements (e.g., replacing "identify" with "locate and recall" for Standard 1 and its goals and objectives). In addition, elements of the *Reading Framework for the 2009 National Assessment of Educational Progress* not included in Exhibit 8 but deemed to be important for alignment purposes were

integrated into the alignment framework (e.g., elements of Exhibit 3, "Literary text matrix: Fiction," and Exhibit 4, "Informational text matrix: Argumentation and persuasive text," were integrated, leading to the inclusion, for example, of Objectives 1.2.e, 1.3.d, and 2.1.e to cover organizing structures). Where needed to resolve overlap and/or ambiguities, the content of multiple objectives was eliminated (e.g., omitting the word "credibility" from Objective 3.3.e to avoid overlap with Objective 3.3.c) or combined (e.g., combining the content of "locate or recall definitions, facts, and supporting details" with the content of "locate or recall specific information in text or graphics" to create a single objective of "locate or recall specific information such as definitions, facts, and supporting details in text or graphics" to reduce redundancy). Conversely, where needed to reflect the intent of the full NAEP framework, additional objective content was added (e.g., adding an objective relating to author's purpose to Standard 2, for consistency across standards; adding the word "perspective" from Exhibit 4 to Objective 2.1.b).

The NAEP reading framework document used in this study reflects all modifications as approved by the Governing Board, with input from the study's technical advisor and ACT.

The ACCUPLACER reading specifications consist of five reading content categories representing the range of assessed content. After extensive collaboration with the College Board and the Governing Board, it was determined that, to most effectively facilitate alignment coding, the College Board would supplement these content categories with brief descriptions intended to elucidate the intent of each category. WestEd added alphanumeric coding to the framework corresponding to the standard (e.g., A) level. For the purposes of this report, the ACCUPLACER specifications categories, coupled with the supplemental descriptions, are referred to as the ACCUPLACER framework. The ACCUPLACER framework used in this study is included in Appendix D.

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.

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 item blocks. Each block consists of either one reading passage or a set of two paired passages, with each passage or passage pair followed by 10–12 items The entire 2009 NAEP grade 12 reading item pool—with the exception of 21 items from four vocabulary blocks[7]—was included in this study. The item pool used consists of 131 passage-based items, organized into 13 single- or paired-passage blocks of approximately 10 items each, and includes multiple-choice items (1 point each) and constructed-response items (1 to 4 points each).

[7] Vocabulary block items are not included in the main NAEP scale and, thus, were excluded from this study, as recommended by the Governing Board.

The ACCUPLACER reading 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 reading item pool, it was unfeasible to include all items in this study, even if the College Board had made the entire 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)—for a total of 70 items. The complete item set for both forms was analyzed for alignment to both the ACCUPLACER framework and the NAEP framework. However, for efficiency, the 15 common items were coded just once by analysts (analysts saw all 55 unique items), and the codes for the common items were weighted as double to retain the balance of content and complexity of the two forms (the full 70 items across both forms).

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 goals 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 40 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 of this size 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 40 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:

- standard (based on cognitive target);
- passage text type;
- text category; and
- item type.

Additionally all items associated with each selected passage were used, and the sample corresponded with four item blocks.[8] The Governing Board reviewed and approved this short version NAEP item pool for use in aligning to the NAEP framework.

NAEP-to-ACCUPLACER Alignment
The entire NAEP item pool of 131 items was analyzed for alignment to the ACCUPLACER framework.

ACCUPLACER-to-ACCUPLACER Alignment
Due to the smaller number of items in the ACCUPLACER selected item pool (two forms), the entire ACCUPLACER selected item pool consisting of 55 items (weighted as 70 items to account for common items) was used to align to the ACCUPLACER framework.

ACCUPLACER-to-NAEP Alignment
The entire ACCUPLACER selected item pool of 55 items (weighted as 70 items to account for common items) was used to align to the NAEP framework.

For alignment purposes, within the WAT system, NAEP items were numbered sequentially, beginning with the first item in the first block. Within the WAT system, all ACCUPLACER Form F items were numbered sequentially in the order in which they appear in the test form, beginning with the first item, followed by all unique items from Form G, numbered sequentially in the order in which they appear in the test form.

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.

[8] In the format in which the total NAEP item pool was provided to WestEd, items appeared with their corresponding passages as "item blocks."

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

	NAEP Grade 12 Reading Assessment	ACCUPLACER Reading Comprehension Assessment

Overall Organization	NAEP reading framework is organized by three interacting categories: **type of text, aspects of text, and cognitive targets**.
	Type of Text is addressed under "Types of Reading Passages"
	Aspects of Text:
	• Genres and types of text referring to the idealized norm of a genre
	• Text structures and features (e.g., point of view, cause and effect), referring to the ways ideas are arranged and connected to one another and to the visual and structural elements that support the reader's comprehension of the text
	• Aspect of author's craft (e.g., voice, symbolism), referring to the specific techniques an author chooses to relay the intended message
	Cognitive Targets:
	• Cognitive dimensions applicable to literary and informational text and specific to each text subtype
	• "The mental processes or kinds of thinking that underlie reading comprehension"
	• Represent a progression from Locate/Recall to Integrate/Interpret to Critique/Evaluate
	Items intended to assess all three cognitive targets
		ACCUPLACER reading framework is organized as follows:
		• Reading Comprehension (75%)
		  • Identifying main ideas (15%-30%)
		  • Direct (explicit) statements/ secondary ideas (15%-40%)
		  • Inferences (15%-40%)
		  • Applications (15%-25%)
		• Sentence Relationships (25%)
		  • Two sentences followed by a question about the relationship

Types of Reading Passages	*Literary texts (30%)*
	• *20% Fiction:* e.g., adventure, historical fiction, realistic fiction, folktales/legends/myths/fantasy, satire, parody, allegory, monologue; intact passages or excerpts
	• *5% Literary nonfiction:* e.g., personal essay, autobiographical/biographical, sketches, speech, character sketches, memoir, classical essay; intact passages or excerpts
	• *5% Poetry:* e.g., narrative poem, free verse, lyrical poem, humorous poem, ode, song, epic, sonnet, elegy; intact poems or excerpts
	Informational texts (70%)
	• *30% Exposition:* e.g., essay, literary
		Informational texts (100%)
		Framework appears to be intended for informational text.
		• Reading comprehension passages can be classified according to the kind of information processing required including explicit statements related to the main idea, explicit statements related to a secondary idea, application, and inference (The College Board, 2010b) (75%):
		  • Identify main idea (15%-30%)
		  • Comprehend specific, explicit information from the passage (15%-40%)
		  • Comprehend details and ideas that are conveyed implicitly in a passage

	NAEP Grade 12 Reading Assessment	ACCUPLACER Reading Comprehension Assessment
	analysis; intact passages or excerpts • *30% Argumentation or persuasive text:* e.g., informational trade book, journal, speech, persuasive essay, letter to the editor, argumentative essay, editorial, historical account, position paper (brochure, campaign literature, advertisement, etc.) • *10% Procedural texts and documents:* e.g., graphics and other information embedded in text, as well as stand-alone documents like applications, manuals, product support materials, and contracts	(inference) (15%-40%) • Understand how the author uses language to achieve his/her purpose in addressing his/her audience (application) (15%-25%) • Sentence relationship passages consist of two sentences followed by a question about the relationship between the sentences (25%)
Characteristics of Reading Passages	As described in the specifications (National Assessment Governing Board, 2009d), NAEP passages provide highly specific criteria for each genre for the selection of reading passages to be used on the test, including (as paraphrased in the specifications): • Well organized, sufficient elaboration of new concepts, use of graphic features (italics, bold print, signal words and phrases) • High quality • Authentic • Coherent • Grade appropriate • Drawn from a variety of contexts • Engaging • Reflecting our literary heritage, including works from varied historical periods • Reviewed for potential bias and sensitivity issues • Reviewed by the Board prior to item development. For each reading passage, NCES will provide the source, author, publication date, passage length, rationale for minor editing to the passage (if any), and notation of such editing applied to the original passage. NCES will provide information and explanatory material on passages deleted in its fairness review procedures. Systematic efforts are made to ensure that texts selected for inclusion on the NAEP Reading Assessment will be interesting to the widest number of students. Readers	No publicly available information, and none of the information furnished for this study describes characteristics of reading passages.

	NAEP Grade 12 Reading Assessment	ACCUPLACER Reading Comprehension Assessment
	become more engaged in text and consequently comprehend a selection better when they find the material interesting. The goal is to ensure that the best possible stimulus material is included on the NAEP Reading Assessment.	
Length of Reading Passages	• Approximately 500–1,500 words • Intended to "gain the most valid information about students' reading" by using material "as similar as possible to what students actually encounter" in and out of school • Long enough to yield a minimum of "10 distinct items"	• Passages are classified as short (35%-40%) and long (15%-25%) • A review of the assessment materials provided for the study shows that the word count for passages used on the ACCUPLACER forms was approximately 33–106.
Reading Difficulty	Difficulty is determined by several methods of selecting and evaluating passages, and other criteria, including: • Expert judgment • Passage mapping • Vocabulary mapping • At least two research-based readability formulas • Grade 12-appropriate reading level • A "variety of sentence and vocabulary complexity" • A thorough review for potential bias and sensitivity	No publicly available information, and none of the information furnished for this study describes reading difficulty of the ACCUPLACER passages.
Vocabulary-Related Tasks	Tasks are intended to determine whether readers know and understand the meanings of the words that writers use to convey new information or meaning, not to measure readers' ability to learn new terms or words. Vocabulary words convey concepts, ideas, actions, or feelings that the readers most likely know. Vocabulary words to be tested: • Characterize the vocabulary of mature language users and characterize written rather than oral language • Label generally familiar and broadly understood concepts, even though the words themselves may not be familiar to younger learners • Are necessary for understanding at least a local part of the context and are linked to central ideas such that lack of understanding may disrupt comprehension	No publicly available information, and none of the information furnished for this study addressed determining the meaning of vocabulary as used in the context of a passage.

	NAEP Grade 12 Reading Assessment	ACCUPLACER Reading Comprehension Assessment
	• Are found in grade-level reading material • Are used in texts from a variety of content domains Tasks are integrated with the other types of passage-based reading comprehension items. In addition, the NAEP item pool includes 21 vocabulary block items that are not linked to passages. These items are not included in the main NAEP scale score, however.	
Number of Items	Items are distributed across multiple test booklets "using a matrix sampling design" so that not all students taking the assessment will receive the same booklets or items. Each student completes: • Two item blocks consisting of two reading passages: 20–24 items total • 3-6 MCs; 5-8 short CRs; and 1 extended CR item per block • 20–30% of items are intertextual	The computer-adaptive version administers 20 items of two primary types: • A reading passage followed by a question based on the text. Both short and long passages are provided. • Sentence relationships items present two sentences followed by a question about the relationship between these two sentences. The "fixed form" version has 35 items.
Item Types	***3–6 Multiple choice*** • 4 answer options: 1 correct, 3 incorrect ***5–8 Short constructed response*** • 1- or 2-sentence response ***1 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. • 40% multiple choice (1 minute each) • 45% short constructed response (2–3 minutes each) • 15% extended constructed response (5 minutes each) • 60% of total test time on constructed responses	Each test is untimed. On the computer-adaptive format, students can change answers to particular questions before moving on to the next question, but cannot leave a question out or come back to it later to change answers.
Assessment Time	Each student spends approximately 50 minutes (2 blocks at 25 minutes each) taking the NAEP Reading Assessment.	The test is untimed but designed to take less than one hour.
When Given	NAEP assesses and reports grade 12 reading results every four years.	ACCUPLACER administrations are determined by colleges and universities using the placement test.
Testing Population	The 2009 Grade 12 NAEP was administered to: • 48,900 12^{th} grade students in reading in 1500 public schools	ACCUPLACER is administered to: • students who are entering or planning to enter college at the freshman level

	NAEP Grade 12 Reading Assessment	ACCUPLACER Reading Comprehension Assessment
	• 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	
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 passages or items read aloud. For a complete list of accommodations: http://nces.ed.gov/nationsreportcard/about/inclusion.asp#accom_table	ACCUPLACER allows use of: • 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

Item Scoring	The items are scored as:
	• Multiple choice:
	  • Incorrect 0
	  • Correct 1
		The items are scored as correct or incorrect. In the computer-adaptive format, correct or incorrect student response impacts the difficulty of the next item received.

	NAEP Grade 12 Reading Assessment	ACCUPLACER Reading Comprehension Assessment

	• Short constructed response:
	  • Incorrect 0
	  • Partial 1
	  • Correct 2
	• Extended constructed response:
	  • Incorrect 0
	  • Partial 1
	  • Essential 2
	  • Extensive 3
	All constructed-response items are scored using rubrics unique to each item. General principles that apply to these rubrics follow:
	• Rubrics define minimal, partial, satisfactory, and extended responses.
	• Students do not receive credit for incorrect responses.
	• All scoring criteria are text based; students must support statements with information from the reading passage.
	• Partial credit is given for responses that answer a portion of the item but do not provide adequate support from the passage.
	• Student responses are coded to distinguish between blank items and items answered incorrectly.
	• Responses are scored on the basis of the response as it pertains to the item and the passage, not on the quality of writing.
	• As part of the item review, the testing contractor will ensure a match between each item and the accompanying scoring guide.	

Test Scores	**Scaled scores:** Range of 0–500; average scores for groups **Achievement levels:** The numeric scale score range is divided into the following three achievement levels: • **Basic** — This level denotes partial mastery of prerequisite skills and knowledge necessary for proficient work at each grade. • **Proficient** — This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of	**Scaled scores:** Range of 20–120 ACCUPLACER provides results measuring the reading skills of individual students. Test scores are used to give college admissions and placement staff information about the academic readiness of students.

	NAEP Grade 12 Reading Assessment	ACCUPLACER Reading Comprehension Assessment
	such knowledge to real-world situations, and analytical skills appropriate to the subject matter. • **Advanced —** This level signifies superior performance. Test scores and achievement levels are used to report on the performance of groups of grade 12 students nationally. In 2009, 11 states participated in the first pilot for reporting state NAEP results at grade 12.	

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 *Web Alignment Tool* (*WAT) Training Manual* (Webb, 2005) in preparation for that training. 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. The facilitators were also asked to re-familiarize themselves with 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 E.

Pre-Workshop Facilitator and Panelist Materials

In preparation for the NAEP–ACCUPLACER pilot study, the study's lead facilitator developed the comparative analysis to document the similarities and differences between the NAEP reading framework and the ACCUPLACER framework. 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 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 *Reading 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, 2009b), and additional background information on the ACCUPLACER printed from the College Board website (College Board, 2010a). 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 F. Copies of facilitator training materials are provided in Appendix E.

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 G). Following this introduction and overview, the two reading panels relocated to a separate room and received training together 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 Items, Grade 12, 2009 (National Center for Educational Statistics, 2009), and from the ACCUPLACER Sample Questions for Students (The College Board, 2007). These sample items were selected by the reading facilitators to represent a range of item types, DOK levels, and objective alignments, and are included in Appendix G.

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 of materials to and receipt of materials at the Westin Grand hotel
- Meeting room security
- Panelist, facilitator, and observer confidentiality agreement
- Secure management of test materials on-site
- Secure management of WAT reports on-site

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

Item Booklets, Framework Documents, and Anchor Papers

WestEd prepared separate bound item booklets for the NAEP and ACCUPLACER assessments. For NAEP, all 131 items were organized by block and numbered sequentially within each block, with the first 40 items identified for coding to the NAEP framework. 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 ACCUPLACER, items were numbered according to the sequence of the ACCUPLACER Forms F and G: the total 35 Form F items, in order, followed by the 20 unique Form G items, in the order in which they appear on the actual forms. 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 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) for use by each panel 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 C. 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.[9]

Process Questionnaires

- A training questionnaire was administered following panelist training on the first day of the alignment workshop to solicit feedback on the training's effectiveness and to identify

[9] For the final ACCUPLACER assessment framework debrief questionnaire, responses were submitted for Panel 1 only.

areas in which more information might be needed. This questionnaire was administered via the online SurveyMonkey system (SurveyMonkey).[10]

- 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 panelist 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 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 panelist 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 assessment 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 the NAEP and ACCUPLACER items (i.e., assigned item numbers and item weights) and frameworks into the WAT, and created the four requisite WAT studies for each group:

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

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

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.

Reading panels used three Westin hotel meeting rooms throughout the alignment workshop. Because this alignment workshop ran concurrently with the NAEP–ACCUPLACER alignment workshop in mathematics, a meeting room large enough to accommodate all reading and mathematics panelists was used for whole-group training and adjudication sessions. A smaller room, large enough to accommodate both reading panels simultaneously, was used for reading combined-panel training and adjudication sessions; this room was also used by one of the reading panels for single-panel coding sessions. A smaller room was used by the other reading panel for single-panel coding sessions.

Each room was equipped with a printer and eight working stations (seven panelist stations and one facilitator station), each one comprising a laptop, a 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 and a technical advisor), who were free to observe panels at their discretion.

Alignment Procedure Implemented in the Study

This alignment workshop occurred over five consecutive days. A full agenda by day is provided in Appendix F, and 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 reading panels worked independently, with the facilitators meeting regularly to discuss progress and decision rules, and to identify items to be discussed 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 G. 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 combined reading panels moved to a joint reading panel meeting room, where the two reading facilitators provided more detailed training in assigning DOK values to objectives. This initial training included group discussion of reading 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 training and evaluation of process questionnaire (provided in Appendix C).

As specified in the design document developed for this project, through the remainder of the week, each panelist independently performed the alignment tasks described in the following subsections (see the study's design document, provided in Appendix A, for a detailed description of each, and see Appendix F 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 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, 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 reading panels convened to be trained in assigning DOK levels to items and mapping items to the NAEP framework. This training included a review of reading DOK levels and both group and individual coding of sample NAEP and ACCUPLACER assessment items.[11] Once the facilitators deemed the panelists to be sufficiently calibrated in coding for both DOK levels and alignment to objectives, the panelists separated into their individual panel rooms. In each group, the facilitator led the

[11] The project director collaborated with the two reading 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 (College Board, 2007). The facilitators then independently coded and reached consensus on DOK levels and alignment to objectives for each item prior to the commencement of this study.

panelists through the coding of a limited sample set of active NAEP items[12] from the item booklet to ensure understanding of the task and calibration among panelists. As indicated earlier, a subset of 40 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 40-item subset to the NAEP framework. Panelists were instructed to record alignment codes for all 40 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 40 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 ongoing 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 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 seven panelists. When relevant items were identified, the facilitators

[12] 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.

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 (Section III of this report).

Review ACCUPLACER Framework and Assign DOK Levels to Each Objective

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 the 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. Therefore, the next step in this alignment workshop's alignment process was for each panel to independently code the ACCUPLACER objectives for DOK. As previously described for the DOK coding of the NAEP objectives, once coding was complete, the two panels individually adjudicated to achieve within-panel agreement on DOK levels; the facilitators met separately to identify and adjudicate differences between the two groups to achieve cross-panel agreement on DOK levels; and, upon reaching cross-panel agreement, the facilitators communicated these values to their panelists and entered ACCUPLACER objectives' DOK values into the WAT.

Map ACCUPLACER Items to the ACCUPLACER Framework

As with the mapping of NAEP items to the NAEP framework, all of the items from the two ACCUPLACER forms were mapped 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[13] 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 and—upon completion of coding—responding to a paper-based debrief questionnaire. 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 were discrepancies between panels of greater than five percentage points.

[13] 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.

Again, consensus was not required, but any issues were communicated to panelists, who had the opportunity to change any codes. These final alignment codes were used by WestEd staff to determine if differences between the two panels were random and not the result of systematic differences in the application of the proposal 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 40 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 40 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. Following the completion of alignment coding, within- and cross-panel adjudication discussions, and completion of the alignment debrief questionnaire, panelists were asked to complete a final ACCUPLACER assessment framework debrief questionnaire, which solicited panelist opinions regarding the overall alignment of both the NAEP and ACCUPLACER assessments to the ACCUPLACER framework.

Map ACCUPLACER Items to the NAEP Framework

The procedures described earlier were used to map ACCUPLACER items to the NAEP framework objectives. Because all 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 ACCUPLACER items to the NAEP framework was limited to determining alignment to objectives.

Pacing and Schedule Adjustments

Throughout the week, the reading panelists completed some coding activities more quickly than had been estimated. In these cases, WestEd staff, in consultation with the COR and facilitators, adjusted the daily schedule as needed. Schedule adjustments were made based on a number of factors, including 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 a subsequent task), and ensuring that new tasks were started following sufficient break time, so that panelists would be refreshed and ready to code. To that end, it was preferable to have panelists dismissed early, rather than to have them begin a new task late in the afternoon, if possible. All 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 content facilitators and agreed to by both panels. The final list of decision rules used for this alignment workshop follows.

NAEP Reading Framework for Alignment: Decision Rules

1. "Simple inferences" in Standard 1 and its associated objectives will be interpreted as including the understanding of close paraphrase of "explicit information" within or across texts.
2. "Author's purpose" in Objective 1.3.b will be interpreted as referring to **explicit** statements of the author's purpose within or across texts. "Author's purpose" in 2.1.f will be interpreted as referring to the **implicit** purpose of a text.
3. "Organizing structures" in Objective 1.3.d will be interpreted as referring to organizing structures that are **explicitly identified** in texts, through such indicators as the author's use of enumeration ("first, second, third," etc.) or explicit references to a problem and its solution (e.g., "The problem is . . ."), etc.
4. The terms "literary devices or text features" in Objective 2.1.d will be interpreted broadly as including all aspects of author's craft and "text features" represented in Exhibits 3 and 4 in the full NAEP reading framework. See examples below.
    - *Literary Devices/Aspects of Author's Craft:* Exaggeration, figurative language (simile, metaphor, symbolism), imagery, connotation, personification, irony, foreshadowing, flashback, comic relief, and dialogue.
    - *Rhetorical Structures/Author's Craft:* Parallel structure, repetition, quotations, analogy, emotional appeal, paradox, contradictions, sarcasm, and irony.
    - *Text Features:* Titles, headings, charts and graphs, italics, bold text, and illustrations.
5. The term "organizing structures" in Objectives 1.3.d and 2.1.e will be interpreted as referring to the organizational structures represented in Exhibits 3 and 4 (comparison, chronology, cause/effect, description, problem/solution, etc.). These objectives will also be interpreted as referring to an author's organization of a larger unit of text (i.e., a paragraph or whole passage), not to the relationship between two sentences.
6. Objective 2.2.c will be interpreted as including the interpretation of character traits or feelings.
7. "Major ideas" in Objective 2.3.a will be interpreted as including important ideas within a paragraph or portion of a text as well as ideas central to a passage as a whole.
8. For Objective 2.3.b, items may be considered fully aligned if they ask students to "draw conclusions" without also requiring them to "provide supporting information." (Some items may ask for both.)
9. When appropriate, items based on literary nonfiction may be aligned to objectives for "informational texts," or for "literary texts," or for objectives that apply to both literary and informational texts.

ACCUPLACER Reading Framework for Alignment: Decision Rules

1. The ACCUPLACER objectives will be interpreted as *not* including the skill of critiquing or evaluating text.
2. ACCUPLACER C may be interpreted as including items asking students to determine the meaning of a word as used in the context of a passage.
3. The ACCUPLACER objectives will be interpreted as not including items addressing the literary element of theme in fiction and poetry or items addressing the unique literary characteristics of poetry (rhythm, rhyme, meter, verse and stanza, sound devices, etc.).[14]
4. ACCUPLACER objectives may be interpreted as including items based on literary nonfiction.
5. ACCUPLACER D may be interpreted as including items addressing mood, tone, or style or an author's use of language.

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.

[14] Based on the comparative analysis of the NAEP framework and the ACCUPLACER specifications, this decision rule was initially developed by the facilitators as "The ACCUPLACER objectives will be interpreted as not including items addressing literary elements of characterization, plot, setting, theme, and organizing structures in fiction or poetry passages." However, panelists' concerns during the operational study about aligning literature-based NAEP items prompted the decision rule to be revised.

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.

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 typical WAT threshold values. If an assessment's framework calls 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 and ACCUPLACER frameworks; they are described as follows:

> *Reading Level 1*. Level 1 requires students to receive or recite facts or to use simple skills or abilities. Oral reading that does not include analysis of the text, as well as basic comprehension of a text, is included. Items require only a shallow understanding of the text presented and often consist of verbatim recall from text, slight paraphrasing of specific details from the text, or simple understanding of a single word or phrase. Some examples that represent, but do not constitute all of, Level 1 performance are:
>
> - Support ideas by reference to verbatim or only slightly paraphrased details from the text.
> - Use a dictionary to find the meanings of words.

- Recognize figurative language in a reading passage.

Reading Level 2. Level 2 includes the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Inter-sentence analysis of inference is required. Some important concepts are covered, but not in a complex way. Standards and items at this level may include words such as summarize, interpret, infer, classify, organize, collect, display, compare, and determine whether fact or opinion. Literal main ideas are stressed. A Level 2 assessment item may require students to apply skills and concepts that are covered in Level 1. However, items require closer understanding of text, possibly through the item's paraphrasing of both the question and the answer. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Use context cues to identify the meaning of unfamiliar words, phrases, and expressions that could otherwise have multiple meanings.
- Predict a logical outcome based on information in a reading selection.
- Identify and summarize the major events in a narrative.

Reading Level 3. Deep knowledge becomes a greater focus at Level 3. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas. Standards and items at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students' application of prior knowledge. Items may also involve more superficial connections between texts. Some examples that represent, but do not constitute all of, Level 3 performance are:

- Explain or recognize how the author's purpose affects the interpretation of a reading selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

Reading Level 4. Higher-order thinking is central and knowledge is deep at Level 4. The standard or assessment item at this level will probably be an extended activity, with extended time provided for completing it. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking. Students take information from at least one passage of a text and are asked to apply this information to a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among texts. Some examples that represent, but do not constitute all of, Level 4 performance are:

- Analyze and synthesize information from multiple sources.
- Examine and explain alternative perspectives across a variety of sources.

- Describe and illustrate how common themes are found across texts from different cultures. (Webb, 2005, pp. 70–71)

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.

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 both of DOK ratings and of coding of objectives and standards to items. For DOK ratings, interrater agreement is calculated as intraclass correlation and pairwise comparison. 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), and values greater than 0.8 reflect good agreement, while values of 0.7 or higher reflect adequate 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 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 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). 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: 0.9371 Pairwise Comparison: 0.7107 *Objective, Standard* Objective Pairwise Comparison: 0.6498 Standard Pairwise Comparison: 0.8803	*DOK* Intraclass Correlation: 0.9651 Pairwise Comparison: 0.7929 *Objective, Standard* Objective Pairwise Comparison: 0.5711 Standard Pairwise Comparison: 0.9048

Sub-Study	Panel 1	Panel 2
Sub-Study 2: ACCUPLACER to NAEP	*DOK* Intraclass Correlation: 0.9128 Pairwise Comparison: 0.7333 *Objective, Standard* Objective Pairwise Comparison: 0.581 Standard Pairwise Comparison: 0.8424	*DOK* Intraclass Correlation: 0.9079 Pairwise Comparison: 0.729 *Objective, Standard* Objective Pairwise Comparison: 0.6231 Standard Pairwise Comparison: 0.8442
Sub-Study 3: ACCUPLACER to ACCUPLACER	*DOK* Intraclass Correlation: 0.9099 Pairwise Comparison: 0.7299 *Objective, Standard* Objective Pairwise Comparison: 0.7818 Standard Pairwise Comparison: 0.7818	*DOK* Intraclass Correlation: 0.9103 Pairwise Comparison: 0.7333 *Objective, Standard* Objective Pairwise Comparison: 0.7463 Standard Pairwise Comparison: 0.7463
Sub-Study 4: NAEP to ACCUPLACER	*DOK* Intraclass Correlation: 0.9458 Pairwise Comparison: 0.7059 *Objective, Standard* Objective Pairwise Comparison: 0.7903 Standard Pairwise Comparison: 0.7903	*DOK* Intraclass Correlation: 0.9654 Pairwise Comparison: 0.7812 *Objective, Standard* Objective Pairwise Comparison: 0.7816 Standard Pairwise Comparison: 0.7816

Looking across panels, Table 2 shows that interrater agreement (within-panel) values for each panel were comparable. Interrater agreement for DOK (intraclass correlation and pairwise comparison) was good for all sub-studies for both groups. Likewise, standard pairwise comparison values were good for all studies for both groups. For match to objective, objective pairwise comparison values were good or reasonable for all sub-studies. For one panel in each of two sub-studies (NAEP-to-NAEP, Panel 2, and ACCUPLACER-to-NAEP, Panel 1), objective pairwise comparison was just below the "reasonable" range, with 0.5711 and 0.581, respectively. Lower objective pairwise comparison values can result from overlapping or unclear objectives, as well as from multiple coding of items. For both of these sub-studies, agreement at the standard level was very high (0.9048 and 0.8424, respectively). Overall, this high level of interrater agreement warrants confidence in the reliability of each panel's findings and the overall conclusions of the study.

As described in Section II of this report, cross-panel agreement attained was monitored throughout the study, as defined 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 and ACCUPLACER Frameworks

Panelists assigned DOK levels to each objective in the NAEP framework and ACCUPLACER framework. The within-panel DOK ratings were then compared across panels, and the two facilitators reached agreement on the final DOK ratings for each objective, discussing with the combined reading panels as appropriate. Consensus DOK values for the NAEP and ACCUPLACER frameworks are shown in Tables 3 and 4, respectively. DOK ratings were assigned to the 37 NAEP objectives and five ACCUPLACER objectives. These ratings are rolled

up to the goal and standard level in the tables. DOK ratings for each objective can be found in Appendix D.

Table 3. DOK Findings for the NAEP 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	1	1 (100%)	-	-	1
1.2	5	5 (100%)	-	-	1
1.3	4	4 (100%)	-	-	1
1 overall	**10**	**10 (100%)**	-	-	**1**
2.1	6	-	1 (17%)	5 (83%)	**2.83**
2.2	5	-	-	5 (100%)	3
2.3	5	-	3 (60%)	2 (40%)	2.4
2.4	1	-	1 (100%)	-	2
2 overall	**17**	**-**	**5 (29%)**	**12 (71%)**	**2.71**
3.1	3	-	-	3 (100%)	3
3.2	3	-	-	3 (100%)	3
3.3	4	-	-	4 (100%)	3
3 Overall	10	-	-	10 (100%)	**3**
ALL	**37**	**10 (27%)**	**5 (14%)**	**22 (59%)**	**2.32**

As shown in Table 3, all objectives in NAEP Standard 1, "Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension," were assigned a DOK level of 1, for an average DOK level of 1. For Standard 2, "Integrate/Interpret: Make complex inferences within and across texts," objectives were assigned DOK ratings of 2 or 3, for an average DOK level of 2.71. For Standard 3, "Critique/Evaluate: Consider text(s) critically," all objectives were assigned a DOK level of 3, and the average DOK level was 3. Overall, the average DOK level of all NAEP objectives was 2.32. Across all standards and the 37 objectives, the distribution of DOK levels was 27% (10) at Level 1, 14% (5) at Level 2, and 59% (22) at Level 3.

The DOK levels of the five ACCUPLACER objectives are shown in Table 4.

Table 4. DOK Findings for the ACCUPLACER Framework

ACCUPLACER Framework	# of Objectives	# and % of Obj. at DOK 1	# and % of Obj. at DOK 2	# and % of Obj. at DOK 3	Average DOK
A. Identifying main ideas	1	-	1 (100%)	-	2
B. Direct statements/ secondary ideas	1	1 (100%)	-	-	1
C. Inferences	1	-	1 (100%)	-	2
D. Applications	1	-	-	1 (100%)	3
E. Sentence relationships	1	-	1 (100%)	-	2
ALL	**5**	**1 (20%)**	**3 (60%)**	**1 (20%)**	**2**

As shown in Table 4, ACCUPLACER Objectives A, "Identifying main ideas," C, "Inferences," and E, "Sentence relationships," were assigned a DOK level of 2. Objective B, "Direct statements/secondary ideas," was assigned a DOK level of 1, and D, "Applications," was assigned a DOK level of 3. Overall, the average DOK level of all ACCUPLACER objectives was 2. Across the five objectives, the distribution of DOK levels was 20% (1) at Level 1, 60% (3) at Level 2, and 20% (1) at Level 3.

Comparing Tables 3 and 4, the objectives in the NAEP framework were found to have an average DOK level of 2.32, compared with an average of 2 in the ACCUPLACER framework. In terms of emphasis of DOK, NAEP has 27% of objectives at DOK Level 1 and ACCUPLACER has 20% at DOK Level 1. NAEP has 14% at DOK Level 2, compared with ACCUPLACER's 60% at DOK Level 2. NAEP has 59% of objectives at DOK Level 3, while ACCUPLACER has 20% at DOK Level 3. In comparing these percentages, however, it is important to consider that NAEP had 37 objectives that could be distributed across the four DOK levels, compared to the 5 ACCUPLACER objectives.

DOK Levels of the Test Items

Panelists assigned each item a DOK rating, independent of any content alignment. 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 40 items used for the NAEP-to-NAEP study were 2.26 for Panel 1 and 2.23 for Panel 2. The average DOK levels of the NAEP items in the complete set of 131 items used for the NAEP-to-ACCUPLACER study were 2.09 for Panel 1 and 2.16 for Panel 2. Thus, the sample appeared representative of the complete pool in terms of DOK. The average DOK levels for the ACCUPLACER items were 1.90 for Panel 1 and 1.93 for Panel 2. The comparison of the DOK levels of the test items with the DOK levels of the objectives they assess is addressed in the depth-of-knowledge consistency analyses later in this section.

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's. 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 surveys 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 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, reviewers evaluated the alignment between the NAEP items and the NAEP framework. A short-form sample of 40 items, corresponding to four passage blocks, was analyzed. The results of Sub-Study 1 are presented in Tables 5–9.

Table 5 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 assigned 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 5. Codability of Items as Determined by Items Rated Uncodable by Seven Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 40

	Panel 1	**Panel 2**
Codable items	40	40
Uncodable items	0	0
Total assessment items	40	40

As shown in Table 5, all 40 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 6 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 6. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Seven Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 40

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	41.14	100	40.43	100
Uncodable	0.00	0	0.00	0
Total	41.14		40.43	

For the 40 items, the range of mean hits across all panelists was 40.43 to 41.14. No uncodable ratings were assigned. These numbers exceed 40 because some items were coded to multiple objectives by one or more panelists.

Table 7 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 7. Categorical Concurrence between Standards and Assessment as Rated by Seven Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 40

	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. Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension	5.57	14	14	9.29	23	23
2. Integrate/Interpret: Make complex inferences within and across texts	31.29	76	76	26.86	66	66
3. Critique/Evaluate: Consider text(s) critically	4.29	10	10	4.29	11	11
Total	41.14	100	100	40.43	100	100

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

Of the three standards, Standard 2 ("Integrate/Interpret: Make complex inferences within and across texts") received the majority of mean hits in both panels (31.29 and 26.86, respectively), making up 66% (Panel 2) and 76% (Panel 1) of the item set. Standard 1 ("Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension") had 5.57 mean hits in Panel 1 and 9.29 in Panel 2, for 14% and 23%, respectively. Standard 3 ("Critique/Evaluate: Consider text[s] critically") received the fewest mean hits, 4.29 in each panel, or 10% (Panel 1) and 11% (Panel 2) of the total hits.

Compared to Panel 1, Panel 2 aligned a higher percentage of items to Standard 1 (23% compared to 14%) and a lower percentage of items to Standard 2 (66% compared to 76%). This discrepancy of approximately four mean hits reflects a slight difference in the two panels' interpretation of the language of Standard 1, particularly in regard to "making simple inferences as needed for literal comprehension." In general, Panel 1 interpreted "simple inferences" more narrowly, assigning items that required anything more than very small, obvious inferences to Standard 2. Panel 2 tended to be slightly broader in their interpretation of "simple inferences," assigning about 10% (4) more items to Standard 1 than did Panel 1. This difference in interpretation was identified during cross-panel adjudication and determined to be minor.

Reporting categorical concurrence in terms of mean hits and percentage of hits at a finer grain size, Table 8 displays the numbers and percentages of mean hits to objectives. Percentages for this table are reported as the percentage of total hits.

Table 8. Number and Percentage of Mean Hits to Objectives as Rated by Seven Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 40

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
1	1.1	1.1.a	4.43	11	5.86	14
	1.2	1.2.a	0.43	1	0	0
		1.2.b	0.14	0	0.29	1
		1.2.c	0	0	0	0
		1.2.d	0	0	0	0
		1.2.e	0	0	0	0
	1.3	1.3.a	0.29	1	2.57	6
		1.3.b	0.14	0	0	0
		1.3.c	0.14	0	0.57	1
		1.3.d	0	0	0	0
2	2.1	2.1.a	0.14	0	0.14	0
		2.1.b	1.57	4	0.43	1
		2.1.c	1.00	2	0.14	0
		2.1.d	5.14	13	3.57	9
		2.1.e	0.29	1	0.29	1
		2.1.f	0.86	2	0.57	1
	2.2	2.2.a	1.00	2	1.14	3
		2.2.b	1.29	3	0.86	2
		2.2.c	2.43	6	2.71	7
		2.2.d	0.86	2	3.57	9
		2.2.e	0	0	0	0
	2.3	2.3.a	1.14	3	1.00	2
		2.3.b	6.71	16	3.00	7
		2.3.c	0.43	1	0.86	2
		2.3.d	0.14	0	0.14	0
		2.3.e	0	0	0.29	1
	2.4	2.4.a	8.29	20	8.14	20

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
3	3.1	3.1.a	1.57	4	1.00	2
		3.1.b	0.57	1	0	0
		3.1.c	0	0	0.14	0
	3.2	3.2.a	0	0	0	0
		3.2.b	0.14	0	1.00	2
		3.2.c	0	0	0	0
	3.3	3.3.a	0	0	0.14	0
		3.3.b	2.00	5	1.57	4
		3.3.c	0	0	0.43	1
		3.3.d	0	0	0	0

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

As shown in Table 8, the following four objectives had the greatest number of mean hits (over five mean hits):

- 1.1.a. “Locate or recall specific information such as definitions, facts, and supporting details in text or graphics”
- 2.1.d. “Describe or analyze how an author uses literary devices or text features to convey meaning”
- 2.3.b. “Draw conclusions and provide supporting information”
- 2.4 a. “Determine word meaning as used in context”

Of these four objectives, Objective 2.4.a received the greatest number of hits and the highest percentage (20%) of total hits of all objectives. Objective 2.4.a is the only objective addressing vocabulary, and one would expect to see all (or nearly all) vocabulary items match that objective. In contrast, there are 36 total objectives to which reading comprehension items could potentially align: 13 specific to literary text, 13 specific to informational text, and 10 applicable to both literary and informational text. Two of the objectives receiving the most hits, 1.1.a and 2.1.d, apply to both literary and informational text (giving them greater potential range of application to items across text types); both are also relatively broad in scope. Objective 2.3.b is one of the most broadly stated (most general) of all objectives for informational texts; it received the second-highest percentage of total alignments, after Objective 2.4.a. Objective 2.3.b is also the only one of the four that is specific to one text type (informational). Decision Rule 9 expanded the interpretation of goal 2.3 to allow for items associated with literary nonfiction because of the overlap with informational text in structure and comprehension strategies applied. The item pool contained one paired passage block containing two literary non-fiction passages; panelists coded 3–5 literary nonfiction items to this objective.

The following objectives received no hits on the NAEP short form:

- 1.2.c. “Locate or recall setting”
- 1.2.d. “Locate or recall figurative language”

- 1.2.e. “Locate or recall organizing structures of literary texts, such as verse or stanza in poetry or description, chronology, comparison, etc., in literary non-fiction”
- 1.3.d. “Locate or recall organizing structures of texts, such as comparison/contrast, problem/solution, enumeration, etc.”
- 2.2.e. “Explain how rhythm, rhyme, sound, or form in poetry contribute to meaning”
- 3.2.a. “Evaluate the role of literary devices in conveying meaning”
- 3.2.c. “Evaluate a character’s conflict, motivations, and decisions”
- 3.3.d. “Judge the coherence or logic of an argument”

Among the objectives receiving no hits, two objectives for Standard 1, 1.2.e and 1.3.d, were objectives about which panelists had raised questions during the pilot study. More specifically, panelists in that study observed that, in most cases, organizing structures of texts are not explicit but have to be inferred from evidence in the text; they had difficulty reconciling this with the activity of locating or recalling explicit content in Standard 1. Following the pilot study, WestEd received clarification about the intent of these objectives from representatives of the Governing Board, and in the operational study, panelists were instructed to interpret these objectives as applying only to texts in which the organizing structure is explicitly identified (for example, by numbering). This principle was articulated in Decision Rule 3 for the NAEP Framework.

Objective 3.2.c, “Evaluate a character’s conflict, motivations, and decisions,” also received no hits. During training and discussion, panelists discussed whether the word “evaluate” was intended to imply judging the ethics or integrity of “a character’s conflict, motivations, and decisions” (in effect, judging the character) or evaluating the literary quality of an author’s portrayal of “a character’s conflict, motivations, and decisions” (judging the author’s craft). However, it appears that panelists did not find any items matching either interpretation of this objective.

In addition to the objectives identified above as having no hits, there were a number of objectives receiving a mean hit value of less than 1.0. A mean hit value of less than 1.0 would indicate that, while the objective was assigned by at least one panelist to an item or items, the objective received fewer than seven total hits across all items and panelists.

Table 9 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 9. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Seven Reviewers per Panel—NAEP Items (Short Version) to NAEP Framework
Assessment items = 40

	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. Locate/Recall	5.57**	9.29	100	100	21**	26**	0.71	0.77
2. Integrate/Interpret	31.29	26.86	65	80	64	59	0.62*	0.68*
3. Critique/Evaluate	4.29**	4.29**	100	100	26**	29**	0.88	0.85

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 40 NAEP assessment items analyzed in the short version, all (40) were found to match objectives, or have "hits." The majority of items were coded to Standard 2, "Integrate/Interpret: Make complex inferences within and across texts." Using the typical WAT threshold value of 6 mean hits, categorical concurrence was met for Standard 2, with 26.86 (Panel 2) and 31.29 (Panel 1) mean hits to the standard. Categorical concurrence was not met for Standard 3, with four mean hits to the standard. The means of the hits assigned to Standard 1 by each panel were 5.57 and 9.29, respectively. However, it is important to note that given the distribution of alignment across the standards, there would be a greater chance of meeting this numerical threshold in the full item pool.

Depth-of-knowledge consistency was met for all three standards, with between 65% 100% of the items at or above the DOK level of the standard to which they aligned. Still well over the threshold value, the depth-of-knowledge consistency for Standard 2 (65% and 80%), compared to that of Standards 1 and 3 (100% for both panels), most likely reflects the greater range of depth of knowledge levels implicit in the language of this standard. The activity of integrating and interpreting texts is notably broad in scope, with potential applications ranging from the interpretation of particular words or phrases in context (typically DOK Level 2) to the summary of main ideas (typically DOK Level 2) to the analysis of theme (DOK Level 3) or of the logical connections across texts (DOK Level 3). In contrast, the language of Standard 1 ("Locate or recall textually explicit information") closely parallels that defining DOK Level 1 ("items . . . often consist of verbatim recall from text"). Similarly, the language of Standard 3 ("Critique/Evaluate: Consider text(s) critically") clearly links it to DOK Level 3 ("Students are encouraged to go beyond the text").

The findings for depth-of-knowledge consistency are also consistent with the DOK levels that panelists assigned to objectives within each standard, presented in Table 3. While 100% of the objectives for Standards 1 and 3 received the same DOK ratings as their corresponding standard (all DOK Level 1 for Standard 1 objectives, and all DOK Level 3 for Standard 3 objectives), objectives for Standard 2 ranged in DOK ratings from Level 2 to Level 3.

Range of knowledge was met for Standard 2, with 59% (Panel 2) and 64% (Panel 1) of the objectives in this standard covered by NAEP items. Range of knowledge was not met for Standards 1 and 3, with only 21%–29% of the objectives in these standards covered (across the two panels). For Standard 1, the majority of aligned items targeted Objective 1.1.a, "Locate or recall specific information such as definitions, facts, and supporting details in text or graphics." For Standard 3, the majority of aligned items targeted either Objective 3.1.a, "Judge the author's craft and technique," or Objective 3.3.b, "Evaluate the strength and quality of evidence used by the author to support his or her position."

Balance of representation was met for Standards 1 and 3 and was weakly met for Standard 2. For Standard 2, while a majority of objectives received hits, the numbers of hits per objective varied considerably. For example, Objective 2.4.a ("Determine word meaning as used in context") received approximately eight mean hits (8.29 and 8.14), while Objective 2.2.a ("Interpret mood, tone, or voice") received just one (1.00 and 1.14). For Standards 1 and 3, while fewer objectives received hits, the distribution of hits among those objectives was more even. For Standard 3, for example, the difference in number of hits per objective averaged approximately 0.5 hits.

Sub-Study 2—ACCUPLACER Items to NAEP Framework

In Sub-Study 2, reviewers evaluated the alignment between the ACCUPLACER items and NAEP framework. Two 35-item forms of the ACCUPLACER assessment were aligned with the NAEP framework, for a total of 70 items. Of these 70 items, 15 were common to both forms, for a total of 55 unique items. The results of Sub-Study 2 are presented in Tables 10–14.

Table 10 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 assigned it an objective. For an item to be uncodable, all reviewers must have rated it uncodable, that is, not aligned to any objective in the framework.

Table 10. Codability of Items as Determined by Items Rated Uncodable by Seven Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 55 (weighted as 70)[15]

	Panel 1	**Panel 2**
Codable items	70	70
Uncodable items	0	0
Total assessment items	70	70

As shown in Table 10, all 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 11 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 11. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Seven Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 55 (weighted as 70)

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	70.00	100	70.00	100
Uncodable	0.00	0	0.00	0
Total	70.00		70.00	

For the ACCUPLACER items, both panels had 70 mean hits. All items were found to align to an objective. No uncodable ratings were assigned.

[15] Each form consists of 35 items—20 variable items and 15 common items—for a total of 70 items. The complete item set for both forms was analyzed. For efficiency, the 15 common items were coded just once by analysts (analysts saw all 55 unique items), and the codes for the common items were weighted as double to retain the balance of content and complexity of the two forms (the full 70 items across both forms).

Table 12 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 12. Categorical Concurrence between Standards and Assessment as Rated by Seven Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 55 (weighted as 70)

	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 - Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension	16.57	24	24	15.71	22	22
2 - Integrate/Interpret: Make complex inferences within and across texts	53.43	76	76	50.57	72	72
3 - Critique/Evaluate: Consider text(s) critically	0.00	0	0	3.71	5	5
Total	70.00	100	100	69.99	100	100

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

Of the three standards, Standard 2 (“Integrate/Interpret”) received the most hits in both panels (53.43 and 50.57, respectively), making up 72% (Panel 2) and 76% (Panel 1) of the item set, results very similar to those for the NAEP items. Standard 1 (“Locate/Recall”) received 16.57 and 15.71 mean hits from the two panels, or 24% and 22% of the total, respectively. Only one panel assigned hits (3.71 or 5%) to Standard 3 (“Critique/Evaluate”), with most of those items assigned to Objective 3.1.b (“Analyze, critique, or evaluate the author’s perspective or point of view”). Panel 1 assigned those same items to Standard 2, primarily to Objective 2.3.b (“Draw conclusions and provide supporting information”). Differences between the two panels regarding the alignment of these items to Standard 3 were identified during the adjudication process and discussed with the panelists. Panel 1 did not find the items to require the critical perspective called for in Standard 3, while Panel 2 saw them as requiring students to “analyze” the “author’s perspective,” a skill covered in Objective 3.1.b. Following the discussion, panelists were given the opportunity to change their ratings based on the conversation. However, the difference in interpretation persisted. Aside from this one issue, the findings of the two panels were very consistent.

Table 13. Number and Percentage of Mean Hits to Objectives as Rated by Seven Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 55 (weighted as 70)

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
1	1.1	1.1.a	13.86	20	10.57	15
	1.2	1.2.a	0	0	0	0
		1.2.b	0	0	0	0
		1.2.c	0	0	0	0
		1.2.d	0	0	0	0
		1.2.e	0	0	0	0
	1.3	1.3.a	1.14	2	3.43	5
		1.3.b	0	0	0	0
		1.3.c	1.43	2	1.43	2
		1.3.d	0.14	0	0.29	0
2	2.1	2.1.a	4.71	7	7.29	10
		2.1.b	17.29	25	13.71	20
		2.1.c	0.71	1	3.43	5
		2.1.d	2	3	3.29	5
		2.1.e	0	0	0.57	1
		2.1.f	1	1	1.14	2
	2.2	2.2.a	0.57	1	1.71	2
		2.2.b	0	0	0	0
		2.2.c	0.14	0	0.14	0
		2.2.d	0	0	0	0
		2.2.e	0	0	0	0
	2.3	2.3.a	5.14	7	6.86	10
		2.3.b	19.71	28	7.14	10
		2.3.c	0.29	0	1.29	2
		2.3.d	0	0	0	0
		2.3.e	0	0	0	0
	2.4	2.4.a	1.86	3	4	6

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
3	3.1	3.1.a	0	0	0.29	0
		3.1.b	0	0	3.14	4
		3.1.c	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
	3.3	3.3.a	0	0	0	0
		3.3.b	0	0	0	0
		3.3.c	0	0	0	0
		3.3.d	0	0	0.29	0

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

As shown in Table 13, the following five objectives had the greatest number of mean hits (over five):

- 1.1.a. “Locate or recall specific information such as definitions, facts, and supporting details in text or graphics”
- 2.1.a. “Describe problem and solution, or cause and effect”
- 2.1.b. “Compare or connect ideas, perspectives, problems, or situations”
- 2.3.a. “Summarize major ideas”
- 2.3.b. “Draw conclusions and provide supporting information”

Of these five objectives, Objective 2.1.b received the greatest number of hits across both panels (17.29 and 13.71); this is one of the broader NAEP objectives, with clear parallels to content in ACCUPLACER C (understand “connections between ideas”) and D (“applying information within the passage . . . to situations outside the passage”). In addition, panelists in both groups assigned some of the sentence relationships items (those typically asking about the logical relationship between two sentences [generalization/example, claim/support, cause/effect, etc.]) to this objective. Objective 2.3.b received the highest number of hits in Panel 1 (19.71) but a considerably lower number (7.14) in Panel 2. Most of the variations between panels were in hits to objectives within Goals 2.1 (both literary and informational text) and 2.3 (informational text only), including objectives with some degree of overlap, such as Objectives 2.1.a, 2.1.b, and 2.3.b. For example, an item asking about cause and effect (2.1.a) could also require drawing a conclusion (2.3.b), so an item could reasonably be assigned to either objective. Some of these variations also likely reflect different distributions of ACCUPLACER E item alignments among these objectives.

Three of the five objectives with over five hits (1.1.a, 2.1.a, and 2.1.b) apply to both literary and informational text, while the other two (2.3.a and 2.3.b) apply only to informational text. This is not surprising given ACCUPLACER’s emphasis on informational texts. In addition, the greater

numbers of hits to these NAEP objectives most likely reflect the overlap between their content and that of several ACCUPLACER objectives. For example, Objectives 2.1.a, 2.1.b, and 2.3.b cover some of the same content as ACCUPLACER C ("Inferences," measuring the "ability to comprehend details and ideas that are conveyed implicitly in a passage, and to understand connections and implications").

NAEP Objective 1.1.a also has considerable overlap with ACCUPLACER B ("Direct statements/secondary ideas," measuring the "ability to comprehend specific, explicit information from the passage and involve the skills of locating information and recognizing and comprehending key details . . ."), while NAEP Objective 2.3.a is closely related to ACCUPLACER A ("Identifying main ideas," referring to the "ability to distinguish the main idea of a passage from supporting ideas OR determine the central focus of a passage even when it is not explicitly stated in the passage").

Small variations across panels in percentages of hits to specific objectives are not uncommon. Also, as this is an analysis of an assessment to a different framework (i.e., not its own), some of the variation may reflect the overlap between some of the NAEP objectives, particularly between some of the broader objectives, such as Objective 2.1.b ("Compare or connect ideas") or Objective 2.3.b ("Draw conclusions"), and some of the more specific ones, such as Objectives 2.1.a ("Describe problem and solution, or cause and effect"), 2.1.c ("Determine unstated assumptions in an argument"), or 2.3.c ("Find evidence in support of an argument"). As can be seen in Table 13, most of the variations across panels were in hits distributed to these related objectives for Goals 2.1 and 2.3.

Only two of the 13 NAEP objectives specific to literary text received any hits. Objective 2.2.a ("Interpret mood, tone, or voice") received 0.57 and 1.71 hits from the two panels, while Objective 2.2.c ("Interpret a character's conflicts, motivations, and decisions") received 0.14 hits from both panels. The low number of hits to literary objectives reflects the emphasis on informational passages in the ACCUPLACER test; panelists did not find ACCUPLACER reading items based on fiction or poetry, but they did find a few items based on literary non-fiction.

As noted previously, only Panel 2 had a small number of hits to objectives for Standard 3; Objective 3.1.b received 3.14 hits from Panel 2, and Objectives 3.1.a and 3.3.d each received 0.29 hits. The low number of hits to objectives in Standard 3 is consistent with key differences in the frameworks for the two assessments (discussed in detail in the NAEP–ACCUPLACER Interim Report). The ACCUPLACER reading passages are considerably shorter than those in NAEP, with one item per passage, in multiple-choice format only. The primary verbs used in the ACCUPLACER objectives are "comprehend," "understand," "recognize," and "distinguish." No ACCUPLACER objectives require students to analyze, critique, evaluate, or judge.

The following objectives received no hits by the ACCUPLACER assessment:

- 1.2.a. "Locate or recall character traits"
- 1.2.b. "Locate or recall sequence of events or actions"
- 1.2.c. "Locate or recall setting"
- 1.2.d. "Locate or recall figurative language"

- 1.2.e. “Locate or recall organizing structures of literary texts, such as verse or stanza in poetry or description, chronology, comparison, etc., in literary nonfiction”
- 1.3.b.”Locate or recall author’s purpose”
- 2.2.b. “Integrate ideas to determine theme”
- 2.2.d. “Examine relations between or among theme, setting, plot, or characters”
- 2.2.e. “Explain how rhythm, rhyme, sound, or form in poetry contribute to meaning”
- 2.3.d. “Distinguish facts from opinions”
- 2.3.e. “Determine the importance of information within and across texts”
- 3.1.c. “Take different perspectives in relation to a text”
- 3.2.a. “Evaluate the role of literary devices in conveying meaning”
- 3.2.b. “Determine the degree to which literary devices enhance a literary work”
- 3.2.c. “Evaluate a character’s conflict, motivations, and decisions”
- 3.3.a. “Evaluate the way the author selects language to influence readers”
- 3.3.b. “Evaluate the strength and quality of evidence used by the author to support his or her position”
- 3.3.c. “Determine the quality of counterarguments within and across texts”

As previously observed, most objectives specific to literary text and most objectives for Standard 3 received no hits. As described in Sub-Study 1, there were also objectives receiving fewer than one mean hit, indicating that few panelists assigned the objective to an item.

Table 14 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 14. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Seven Reviewers per Panel—ACCUPLACER Items to NAEP Framework
Assessment items = 55 (weighted as 70)

	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 - Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension	16.57	15.71	100	100	27**	31**	0.55**	0.65*
2 - Integrate/Interpret: Make complex inferences within and across texts	53.43	50.57	39**	49*	45*	58	0.55**	0.71
3 - Critique/Evaluate: Consider text(s) critically	0**	3.71**	0**	79	0**	11**	0**	0.84

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 70 ACCUPLACER assessment items analyzed, all (70) were found to match or "hit" objectives. The majority of items were coded to Standard 2, "Integrate/Interpret: Make complex inferences within and across texts." Using typical WAT threshold values, categorical concurrence was met for Standard 2, with 50.57 (Panel 2) and 53.43 (Panel 1) mean hits to the standard. Categorical concurrence was also met for Standard 1, with 15.71 (Panel 2) and 16.57 (Panel 1) mean hits to the standard. Categorical concurrence was not met for Standard 3, with fewer than four mean hits from one panel to the standard.

Across both panels, depth-of-knowledge consistency was met only for Standard 1, with 100% of the items rated at or above the DOK level of the standard. Standard 2 showed a weak depth-of-knowledge consistency in Panel 2 (49%) and did not meet the threshold for depth-of-knowledge consistency in Panel 1 (39%). Both panels found that a majority of items aligned to Standard 2 had a lower DOK level than that of the standard. The difference in depth-of-knowledge consistency results for Standard 2 was identified in cross-panel adjudication as the result of items that could reasonably be interpreted as either DOK Level 2 or Level 3. For Standard 3, the difference was found to be systematic, in that, for the 2–3 items, as described earlier, there was a difference in the two panels' interpretation of the extent of analysis required by the item. Therefore, while this standard meets the criterion in Panel 2, there were no items assigned in Panel 1.

Range of knowledge was weakly met or met for Standard 2, with hits to 45% and 58% of objectives for that standard. Range of knowledge was not met for Standard 1, with 27% and 31%

of its objectives hit, or for Standard 3, with 11% of objectives hit in Panel 2 only. The limited range of knowledge found in the ACCUPLACER items is also consistent with differences in the frameworks of the two tests: the ACCUPLACER framework includes just five broad objectives while the NAEP framework includes 37 more specific objectives. In addition, the ACCUPLACER objectives do not specifically address literary text and do not refer to the skills of critiquing, evaluating, or judging the quality or effectiveness of a text or an author's craft; therefore, one would expect ACCUPLACER to cover a narrower range of skills.

The results of the two panels differed on balance of representation. In Panel 1, balance of representation was not met for any standard, while in Panel 2, balance of representation was met for Standards 2 and 3 and weakly met for Standard 1. Compared to Panel 1, then, Panel 2 distributed hits more evenly among the objectives receiving hits for each standard. In general, Panel 1 assigned more hits to the broader (more general) NAEP objectives, such as Objectives 2.1.b ("Compare or connect ideas") and 2.3.b ("Draw conclusions"), while Panel 2 assigned more hits to some of the more specific NAEP objectives, such as Objectives 2.1.c ("Determine unstated assumptions in an argument") and 2.1.e ("Describe or analyze how an author uses organizing structures to convey meaning"). For the most part, the variation in distribution of hits to objectives occurred between related objectives within NAEP Goals 2.1 and 2.3. As noted earlier, however, Panel 2 also assigned some items to Standard 3, Objective 3.1.b, while Panel 1 assigned those items to Objective 2.1.b.

Sub-Study 3—ACCUPLACER Items to ACCUPLACER Framework

In Sub-Study 3, reviewers evaluated the alignment between the ACCUPLACER items and the ACCUPLACER framework. Two 35-item forms of the ACCUPLACER Reading test were analyzed, for a total of 70 items. Of these 70 items, 15 were common to both forms, for a total of 55 unique items. The results of Sub-Study 3 are presented in Tables 15–18.

Table 15 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 assigned it an objective. For an item to be uncodable, all reviewers must have rated it uncodable, that is, not aligned to any objective.

Table 15. Codability of Items as Determined by Items Rated Uncodable by Seven Reviewers per Panel—ACCUPLACER Items to ACCUPLACER Framework
Assessment items = 55 (weighted as 70)[16]

	Panel 1	**Panel 2**
Codable items	70	70
Uncodable items	0	0
Total assessment items	70	70

As shown in Table 15, all 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 16 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 16. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Seven Reviewers per Panel—ACCUPLACER Items to ACCUPLACER Framework
Assessment items = 55 (weighted as 70)

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	70.00	100	70.00	100
Uncodable	0.00	0	0.00	0
Total	70.00		70.00	

For the ACCUPLACER items, both panels had 70 mean hits. All items were found to align to an objective. No uncodable ratings were assigned.

[16] Each form consists of 35 items—20 variable items and 15 common items—for a total of 70 items. The complete item set for both forms was analyzed. For efficiency, the 15 common items were coded just once by analysts (analysts saw all 55 unique items), and the codes for the common items were weighted as double to retain the balance of content and complexity of the two forms (the full 70 items across both forms).

Table 17 shows the categorical concurrence based on the counts of items that were coded to each of the five objectives 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 17. Categorical Concurrence between Standards and Assessment as Rated by Seven Reviewers per Panel—ACCUPLACER Items to ACCUPLACER Framework
Assessment items = 55 (weighted as 70)

	Panel 1			Panel 2		
Objectives	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
A. Identifying main ideas	9.57	14	14	10.86	16	16
B. Direct statements/ secondary ideas	15.00	21	21	14.29	20	20
C. Inferences	20.71	30	30	18.71	27	27
D. Applications	7.71	11	11	9.00	13	13
E. Sentence relationships	17.00	24	24	17.14	24	24
Total	70.00	100	100	70.00	100	100

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

Of the five objectives, ACCUPLACER C (“Inferences”) received the most hits (18.71 and 20.71) in Panels 2 and 1, respectively, accounting for 27% and 30% of the total items. ACCUPLACER E (“Sentence relationships”) received 17.00 and 17.14 hits (24% of total for each panel), and ACCUPLACER B (“Direct statements/secondary ideas”) received 14.29 and 15 hits (20% and 21%) of the total hits. ACCUPLACER A (“Identifying main ideas”) and ACCUPLACER D (“Applications”) received the fewest hits of the five standards, with 9.57 and 10.86 hits for ACCUPLACER A (14% and 16%) and 7.71 and 9.00 hits for ACCUPLACER D (11% and 13%).

Table 18 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 18. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Seven Reviewers per Panel—ACCUPLACER Items to ACCUPLACER Framework

Assessment items = 55 (weighted as 70)

	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)	
Objectives	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
A. Identifying main ideas	9.57	10.86	92	79	100	100	1	1
B. Direct statements/ secondary ideas	15.00	14.29	100	100	100	100	1	1
C. Inferences	20.71	18.71	100	98	100	100	1	1
D. Applications	7.71	9.00	87	74	100	100	1	1
E. Sentence Relationships	17.00	17.14	92	96	100	100	1	1

All alignment criteria were met by all of the ACCUPLACER objectives. As seen in Table 18, depth-of-knowledge consistency was slightly lower for ACCUPLACER A and D; the language of both of these broadly stated standards suggests a range of DOK levels. ACCUPLACER A, for example, includes “recognition of paraphrase” (typically DOK Level 1) and determining “the central focus of a passage even when it is not explicitly stated” (typically DOK Level 2). ACCUPLACER D also has a broad range, so that some variation in DOK levels would be expected. All objectives met the criteria for depth-of-knowledge consistency and categorical concurrence. The difference of greater than five percentage points for depth-of-knowledge consistency in ACCUPLACER A and ACCUPLACER D was identified during cross-panel adjudication. Facilitators reviewed the discrepant items and found that, while Panel 1 had tended to code them higher than Panel 2, the items could reasonably be coded between the DOK levels.

All five also had 100 percent range of knowledge and balance of representation, but it is important to remember that, with a framework consisting of only one hierarchical level, these criteria can be met with one mean hit. Therefore, range of knowledge and balance of representation are not applicable when referring to the ACCUPLACER framework.

Sub-Study 4—NAEP Items to ACCUPLACER Framework

In Sub-Study 4, reviewers evaluated the alignment between the NAEP items and the ACCUPLACER framework. All 131 NAEP items were analyzed. The results of Sub-Study 4 are presented in Tables 19–22.

Table 19 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 assigned it an objective. For an item to be uncodable, all reviewers must have rated it as uncodable, that is, not aligned to any objective.

Table 19. Codability of Items as Determined by Items Rated Uncodable by Seven Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 131

	Panel 1	**Panel 2**
Codable items	123	121
Uncodable items	8	10
Total assessment items	131	131

As seen in the table, both panels found a small number of NAEP items (8 and 10) to be uncodable to the ACCUPLACER objectives. The specific skills included in the uncodable items are discussed following Table 20.

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 20 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 20. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Seven Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 131

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	114.86	88	112.00	85
Uncodable	16.14	12	19.00	15
Total	131.00		131.00	

While Table 19 shows the number of items that all seven reviewers in each panel rated as codable or uncodable, Table 20 shows the numbers and percentages of mean hits for codable and uncodable items within each panel, counting all hits in a panel, not only those with 100% agreement. As shown in Table 20, there were a number of additional uncodable ratings assigned by each panel, and both panels rated similar percentages of items as codable (85% and 88%) or uncodable (12% and 15%). In fact, as the item-level WAT tables in Appendix M (Tables 12.8 and 12.9) show, a majority of reviewers (5 or more of 7) in both panels rated the same 15 items as uncodable, a very strong degree of agreement across both panels. Reviewers' comments show

two rationales for determining that NAEP items were uncodable to ACCUPLACER objectives: 1) the item calls for evaluation, a skill not covered by any ACCUPLACER objective, or 2) the item asks about the theme (or other specifically literary feature) of a literary work, a skill not addressed by any ACCUPLACER objective.

Table 21 shows the categorical concurrence based on the counts of items that were coded to each of the five objectives 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 21. Categorical Concurrence between Standards and Assessment as Rated by Seven Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 131

	Panel 1			**Panel 2**		
Objectives	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
A. Identifying main ideas	7.86	7	6	7.71	7	6
B. Direct statements/ secondary ideas	29.71	26	23	30.00	27	23
C. Inferences	56.43	49	43	51.43	46	39
D. Applications	20.86	18	16	22.86	20	17
E. Sentence Relationships	0.00	0	0	0.00	0	0
Total	114.86	100	88	112.00	100	85

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

Of the five ACCUPLACER objectives, ACCUPLACER C (“Inferences”) received the most hits (51.43 and 56.43), accounting for 46% and 49% of total hits. This result is not surprising given the breadth of the standard and its overlap with the NAEP framework. ACCUPLACER C is a broadly stated objective and includes making inferences to understand “details and ideas that are conveyed implicitly in a passage” as well as “connections and implications.” At a more general level, its content overlaps with that of a number of NAEP Standard 2 objectives (2.1.a, 2.1.b, 2.1.c, 2.3.b, 2.3.c, and 2.3.d, for example), as well as with the language of Standard 2 itself, “Make complex inferences within and across texts.” In addition, Decision Rule 2 for ACCUPLACER allowed panelists to match NAEP vocabulary-in-context items to ACCUPLACER C, as these items require making inferences from context clues. ACCUPLACER B (“Direct statements/secondary ideas”) had 29.71 and 30 hits (26% and 27%), and ACCUPLACER D (“Applications”) received 20.86 and 22.86 hits, or 18% and 20% of total hits. Both broad standards also include content parallel to that of a number of more specific NAEP objectives. For example, ACCUPLACER B, which includes the ability to understand “specific, explicit information” in a passage, as well as “key details” and “secondary ideas,” has parallels in NAEP Objectives 1.1.a, 1.3.a, and 1.3.c, while ACCUPLACER D includes content related to NAEP Objectives 2.1.b, 2.1.c, and 2.1.d, though at a more general level.

ACCUPLACER A received a smaller number of hits (7.71 and 7.86, or 7%) compared to ACCUPLACER B, C, and D. A more specific standard, ACCUPLACER A is most closely related to NAEP Objectives 2.3.a (“Summarize major ideas”) and 1.3.a (“Locate or recall the topic sentence or main idea”). ACCUPLACER E received no hits; the NAEP framework does

not address the skill of "understanding the relationship between two sentences," and, therefore, no NAEP items address this skill explicitly.

After adjusting for the 12% and 15% uncodable hits, the distribution remains consistent across the objectives, but the percentages reduce proportionally.

Table 22 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 22. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Seven Reviewers per Panel—NAEP Items to ACCUPLACER Framework
Assessment items = 131[17]

	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)	
Objectives	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
A. Identifying Main Ideas	7.86	7.71	79	86	100	100	1	1
B. Direct Statements/ Secondary Ideas	29.71	30	100	100	100	100	1	1
C. Inferences	56.43	51.43	96	99	100	100	1	1
D. Applications	20.86	22.86	86	78	100	100	1	1
E. Sentence Relationships	0**	0**	0**	0**	0**	0**	0**	0**

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 131 NAEP items analyzed, most were found to match ACCUPLACER objectives, although 8 and 10 items were determined to be uncodable by all panelists, and 12% and 15% of ratings were uncodable items. The values in the table above reflect codable items only. As previously noted, ACCUPLACER C ("Inferences") received the most hits, almost double the number of any other standard. All alignment criteria were met for ACCUPLACER A and D, with 100 percent range of knowledge and balance of representation, but it is important to remember that, with a framework consisting of a single hierarchical level, these criteria can be met with one mean hit. Therefore, range of knowledge and balance of representation are not applicable when referring to the ACCUPLACER framework. ACCUPLACER E did not meet any of the alignment criteria; NAEP does not cover this skill. The discrepancies of greater than five percentage points between panels in depth-of-knowledge consistency for ACCUPLACER A and D were identified and discussed during cross-panel adjudication. After reviewing the related items, facilitators determined that the differences were not systematic.

[17] It should be noted that, as shown in Table 20, 12% and 15% of the adjusted total hits for NAEP items were determined by panelists to be uncodable to any ACCUPLACER objective.

IV. Panelists' Evaluations of the Process

This section details the findings from responses to training and process evaluation questionnaires that the seven panelists from each of two panels (a total of 14 panelists) completed at the end of each day of participation. WestEd administered these questionnaires to determine what factors, if any, might impede consistent and reliable alignment coding within and across panels, and 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 23 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 23. Panelist Responses to Day 1 Training and Process Evaluation Questionnaire

	Panel 1 (n=7)				**Panel 2** (n=7)			
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	2	**5**	0	0	2	**5**
Q6. introduce NAEP/ ACCUPLACER?	0	0	2	**5**	0	0	**4**	3
Q7. prepare you to under-stand DOK levels?	0	0	2	**5**	0	0	2	**5**
Q8. prepare you for the consensus process?	0	0	1	**6**	0	0	2	**5**
Q9. prepare you to use the WAT system?	0	0	1	**6**	0	**3**	1	**3**
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	1	2	**4**	0	1	**4**	2
How well did your facilitator…	**Not Well**	**Moderately Well**		**Very Well**	**Not Well**	**Moderately Well**		**Very Well**
Q13. facilitate today's consensus process?	0	0		**7**	0	0		**6***

*n=6 for this question.

As shown in Table 23, 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, prepared them for understanding definitions of DOK levels, and

prepared them for the discussion process that led to agreement on DOK levels for NAEP objectives across the two panels. All but three (43%) panelists reported feeling either adequately or very well prepared to use the WAT system. The three panelists who reported feeling somewhat prepared to use the WAT were in Panel 2. Of these, one stated that s/he would have liked more time to learn the system but felt confident that s/he would gain necessary familiarity; neither of the other two panelists commented further on the WAT training. All panelists were monitored by facilitators and WestEd staff over the course of the workshop to ensure that they could effectively use the WAT, and all were able to manage the system throughout the alignment process.

When asked how comfortable they felt with the process of assigning DOK levels to objectives, the majority of panelists on both panels reported feeling either comfortable or very comfortable. One panelist on each panel reported feeling somewhat comfortable with the process: the panelist from Panel 1 indicated that s/he felt very well prepared for assigning DOK levels to objectives and for the discussion process, and this panelist had no recommendations for improving the process; the panelist from Panel 2 also felt very well prepared for assigning DOK levels to objectives and for the discussion process, and this panelist recommended providing more space on the framework document itself to record notes.

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. When panelists were asked if any additional information would be useful, the two requests made were to receive more practice in alignment and to learn more about WestEd's role as a national assessment service. Recommendations for improving the training and alignment process included allowing more time for calibration, providing the decision rules to panelists earlier in the day, providing more room on the framework document to write ratings, taking time to review the framework as a whole group, and providing pre-coded objectives/items to panelists as part of the training.[18] When asked what aspects of the day went particularly well, 57% (8) of the 14 panelists mentioned the group discussion processes; other responses included the morning's overview session (14%, or 2) and the decision rules (7%, or 1).

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 to have confidence in the Day 1 assignment of DOK levels to objectives and to move into the Day 2 activities. WestEd staff monitored both panels to ensure that all panelists were able to complete the remaining alignment activities, and felt comfortable doing so, 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 assigned DOK values to ACCUPLACER framework objectives. At the end of the day,

[18] Training included coding of sample objectives and items, including discussion of the correct answer or answers most consistent with the criteria. Care was taken not to overly influence panelists through the examples.

panelists were administered a questionnaire that solicited feedback on training for assigning DOK values to items and for aligning items to objectives; it also solicited feedback regarding panelists' comfort with the day's alignment activities. Table 24 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 24. Panelist Responses to Day 2 Training and Evaluation of Process Questionnaire

	Panel 1 (n=7)				**Panel 2** (n=7)			
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	1	1	**5**	1	0	1	**5**
Q5. prepare you for the alignment (coding) process?	0	1	**3**	**3**	0	0	3	**4**
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	0		**7**

Overall, panelists reported feeling prepared to assign DOK levels to items and to code items to objectives. However, one panelist from Panel 2 reported feeling not well trained (expressing a concern that the two panels debated over agreement on an item because the two panels considered different factors when making the decision) and one panelist from Panel 1 reported feeling only somewhat well trained, to assign DOK levels to test item (indicating that the transition from coding DOK of NAEP objectives and items to coding DOK of ACCUPLACER items was a challenge) Proficiency in and comfort level with assigning DOK levels to items were monitored for all panelists in both panels on Day 3; at the end of Day 3, the aforementioned panelist from Panel 2 reported feeling very comfortable with assigning DOK levels to items, while the aforementioned panelist from Panel 1 reported feeling comfortable. No panelist reported feeling not well prepared for the alignment coding process, although the panelist in Panel 1 who reported feeling somewhat prepared for the DOK coding also reported feeling only somewhat prepared for item alignment coding; this panelist indicated that s/he would have liked more time working with the ACCUPLACER framework before beginning alignment. However, by the end of Day 3, this panelist reported feeling comfortable with both processes. This panelist was also the only one of the 14 reading panelists to report that the facilitators did only moderately well in managing the within-panel discussions regarding alignment codes[19], commenting that having more "explicit imperatives (direct instructions)" about what to do (e.g., discuss with panelists, work independently), and when to do it, would have more effectively kept her/him on task.

Recommendations solicited by this Day 2 questionnaire for improving the alignment process or requests for more information included providing more examples and more practice with the ACCUPLACER framework before coding items to the framework. Three panelists commented on the value of within-panel discussions regarding alignment codes, with additional comments relating to reaching agreement more quickly (one panelist) and increasing consistency in criteria

[19] 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 design document.

across panels (one panelist). When asked what activities went particularly well, five panelists indicated the within-panel discussions, two panelists indicated practice coding, and one panelist indicated ACCUPLACER DOK coding. When asked to identify areas in which they felt unprepared, one panelist reported some difficulty in shifting from the NAEP framework to the ACCUPLACER framework, and one referenced a specific coding issue ("drawing conclusions" versus "connections").

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 25 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 25. Panelist Responses to Day 3 Evaluation of Process Questionnaire

	Panel 1 (n=7)				**Panel 2** (n=7)			
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	**4**	3	0	1	1	**5**
Q5. aligning test items to objectives?	0	0	**6**	1	0	1	**4**	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	**7**		0	0	**7**	

All but one panelist felt comfortable or very comfortable with Day 3's alignment activities; the one panelist felt somewhat comfortable. Day 3's activities consisted largely of coding NAEP items to the ACCUPLACER framework, and panelists informally reported that coding an assessment to another assessment's framework was more difficult than coding an assessment to its own framework. Furthermore, as indicated in Section II of this report, a preliminary decision rule—relating to the ACCUPLACER framework—that had been established prior to the study was refined during the alignment coding process to better reflect the experiences of the panelists in coding. One panelist in Panel 2 reported feeling only somewhat comfortable assigning DOK levels and alignment codes; this panelist commented that the decision rule modification may have caused indecision in coding items.

The questionnaire asked panelists to provide recommendations for improving the alignment process, to record requests for more information, and to specify activities for which they felt unprepared. Five of the 14 panelists commented on changing the decision rule during the alignment process—three indicated that it was confusing to change, and two endorsed the change. Another panelist commented about the difficulty in coding NAEP items to the ACCUPLACER framework. When asked what activities went particularly well, panelists expressed appreciation for examples and opportunities to discuss. When asked what could have

been done differently to improve the day's activities, the majority (71%, or 5, of the seven responses to this question) were positive, indicating that the alignment activities went well.

Day 4 Process Evaluation

The remaining alignment activities—primarily coding ACCUPLACER items to the NAEP framework—were conducted on the fourth day of the study. At the end of the day, panelists were administered a process evaluation questionnaire that solicited feedback on these alignment activities. Table 26 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 26. Panelist Responses to Day 4 Process Evaluation Questionnaire

	Panel 1 (n=7)				**Panel 2** (n=7)			
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	3	**4**	0	1	2	**4**
Q5. aligning test items to objectives?	0	1	**5**	1	0	2	**3**	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	1	**6**		0	0	**7**	

Overall, most panelists reported feeling comfortable or very comfortable assigning DOK levels to items, with one panelist in Panel 2 reporting feeling only somewhat comfortable with this process. This panelist also reported feeling somewhat comfortable with aligning test items to objectives, although s/he provided no elaboration on the questionnaire, commenting only that the adjudication process helped clarify and explain NAEP objectives. A second panelist in Panel 2 reported feeling only somewhat comfortable with the item alignment process, commenting that the adjudication process occurred too quickly to thoughtfully consider code changes; one panelist in Panel 1 also reported feeling somewhat comfortable with item alignment, although this panelist made no further comment. One panelist in Panel 1 indicated that the facilitator conducted the within-panel discussion moderately well, without leaving any additional comment; all other panelists reported that the facilitators conducted the day's within-panel discussion process very well. During a debrief that occurred at the end of Day 4, neither WestEd staff nor facilitators identified concerns with the process or with individual panelist codes that would call final alignment codes into question.

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; two panelists requested more decision rules and/or more explicit coding instructions to facilitate coding, and one panelist reported having insufficient time during the within-panel discussion process to think through possible changes to codes. One panelist explicitly reported difficulty aligning across frameworks. Overall, however, responses were largely positive, with panelists again expressing appreciation for opportunities to discuss and adjudicate alignment decisions.

Day 5 End-of-Study Evaluation

On the final day of the study, panelists responded to additional questions about the alignment process, the effectiveness of their panels and facilitators, and study logistics. 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 27 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 27. Panelist Responses to End-of-Study Questionnaire

	Panel 1 (n=7)				**Panel 2** (n=7)			
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	**4**	3	0	0	1	**6**
Q7. for the consensus process?	0	1	2	**4**	0	0	0	**7**
Q8. for the alignment (coding) process?	0	0	**4**	3	0	1	2	**4**
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	0	**4**	3	0	0	1	**6**
Q6. assigning DOK levels to test items?	0	0	**4**	3	0	0	2	**5**
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	1	**3**	**3**	0	2	**5**	0
Q10. on-site training and coding materials?	0	0	0	**7**	0	0	2	**5**
How qualified was your panel…	**Not Quali-fied**	**Some-what Quali-fied**	**Adeq-uately Quali-fied**	**Very Quali-fied**	**Not Quali-fied**	**Some-what Quali-fied**	**Adeq-uately Quali-fied**	**Very Quali-fied**
Q11. to conduct this type of alignment?	0	0	2	**5**	0	0	0	**7**
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?	0	0	1	**6**	0	0	2	**5**

By the end of the week, as evidenced on the end-of-study survey, all panelists across the two panels reported feeling adequately or very well trained in DOK coding, and all but two panelists reported feeling at least adequately trained to participate in within-panel discussions and align items to objectives; neither of these two panelists provided any other comment regarding why they felt only somewhat well prepared by Day 1 training. Over the course of the workshop, the

majority of the 14 panelists reported feeling either comfortable or very comfortable assigning DOK levels and alignment codes; however, on each day's questionnaire, between one and three panelists reported feeling somewhat comfortable with these processes, with one panelist on Day 2 reporting that the training had not trained her/him well to assign DOK levels to the test items. The challenges faced when aligning an assessment to another assessment's framework was frequently raised as a concern. On the end of study questionnaire, however, all panelists reported feeling either comfortable or very comfortable with coding DOK levels of objectives and coding DOK levels of items.

Twelve of the 14 panelists reported that their panels were either adequately qualified or very qualified to conduct the alignment activities, although two panelists on Panel 1 were less positive about their panel's qualifications. When asked what could have been done to improve the composition of the panel, one of these panelists commented that more experts in the field would have enhanced the panel, while the second of these two panelists commented that the panel was thoughtful and thorough and that s/he would not alter its composition; on this open-ended question, four of the remaining five panelists on this panel indicated that the panel was well-balanced and qualified, while one panelist who rated the panel as very qualified suggested that some members might have been more dominant than others. Overall, therefore, panelist perceptions of their own panels were positive, corresponding with perceptions of both facilitators and WestEd staff.

The majority (79%, or 11) of panelists indicated that the information they received in advance of the study was adequately useful or very useful, while all panelists reported that the training and coding materials they received during the alignment workshop were either adequately useful or very useful. Overall, they also responded positively to the WAT, although some commented about minor technical problems (e.g., timing out).

When asked to provide qualitative feedback about their facilitator's effectiveness in managing adjudication procedures, all but one of the 14 panelists reported that facilitators were very effective, providing guidance and engaging the entire group without overpowering the group's decisions. The remaining panelist, from Panel 2, indicated that at times the facilitator was too overt in guiding group discussions, although overall the facilitator was effective and provided adequate time for group processing.

More substantive questions were asked via open-ended questions. Regarding the utility of the alignment criteria in capturing aspects of each assessment (Question 14), panelist responses focused either on the alignment criteria or the frameworks used in the study. Those who discussed the criteria reported them to be useful, in particular for within-assessment alignment. One panelist responded that the criteria were less useful for cross-assessment alignment, with whole categories eliminated.

All but one panelist reported that the alignment process effectively captured content *similarities* between the assessments, with this one panelist commenting that the assessments cover different kinds of knowledge and that it was difficult to align ACCUPLACER items to NAEP's specific objectives. Six of the panelists specifically referenced perceived similarities between the assessments: one panelist reported similarities at the broader category level (e.g., inference, analysis, or recall), with similarities between nuances within each of these categories less

apparent; one panelist reported similarities at the more detailed levels and less at the higher, broader levels in the frameworks; and one panelist reported similarities in depth of knowledge (both assessments largely coded at DOK Level 2) and in a shared emphasis on inference.

Similarly, all panelists reported that the alignment process effectively captured content *differences* between the assessments, although one panelist suggested that having different names and labels for categories across the two frameworks might lead to a perception of greater difference than is warranted, and that while NAEP has deeper reasoning skills than does ACCUPLACER, this difference does not necessarily imply that the content of the two assessments is different. Furthermore, two panelists noted that the process highlighted an underlying difference in specificity between the two assessments (e.g., NAEP is more specific and ACCUPLACER is more abstract), while another panelist observed that ACCUPLACER does not address support for one's position/argument or the evaluation of text, while NAEP does not address sentence-level analysis to the same extent as ACCUPLACER. One panelist raised a concern that differences in genres covered and lengths of passages between the two assessments would not be reported in quantifiable terms, although such differences are significant. Another panelist voiced an opinion that more context for interpreting the criteria (e.g., expectations of the tested population of 12th graders) would have been useful.

The final evaluation survey also included a question about the assessments themselves, and panelists reported similarities and differences related to item type, passages, frameworks, and content assessed in the assessments. The NAEP assessment was seen to address inference tasks, critical thinking, evaluative/analytical writing, rhetorical process, and literary terms more so than the ACCUPLACER assessment, whereas ACCUPLACER had a greater focus on sentence relationships, basic comprehension of more straightforward passages, and a limited range of thinking skills. The two assessments were also seen to differ with respect to genre—with NAEP covering both fiction and informational texts, while ACCUPLACER covers only informational text—and item type—with NAEP including both multiple-choice and constructed-response items, compared with ACCUPLACER's multiple-choice-only format. On the other hand, one panelist indicated that ACCUPLACER offered concise but challenging items, whereas NAEP's items are unnecessarily diverse. As for areas of similarity, one panelist suggested that the two assessments share the same basic level of readability.

When asked about the facilities for the alignment workshop, panelists felt that they were suitable. Table 28 shows panelist responses to this question, by panel. Numbers in bold font represent the highest number of responses for each question, by panel. Only five panelists from Panel 1 completed this questionnaire.

Table 28. Panelist Responses Regarding Adequacy of Facilities

	Panel 1 (n=5)				**Panel 2** (n=7)			
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	0	2	**3**	0	1	2	**4**
Computers and equipment	0	0	0	**5**	0	0	3	**4**
Meals and breaks	0	0	0	**5**	0	0	2	**5**
Sleeping rooms	0	0	0	**5**	0	0	0	**7**

Across both panels, all but one of the panelists who responded felt that the meeting rooms were very suitable or adequately suitable, while one panelist found the meeting rooms somewhat suitable. All panelists who responded felt the computers, equipment, meals, and breaks were either very suitable or adequately suitable for this type of alignment workshop, and all panelists who responded felt the sleeping rooms were very suitable.

V. Summary and Conclusions

Section III reported various indices of alignment for each sub-study individually. This section begins with a summary of the content overlap of content alignment, followed by a summary of alignment of each assessment vis-à-vis the four criteria of the study, and ends with conclusions regarding the alignment of the NAEP and ACCUPLACER assessments.

Summary of Overlap of Content Alignment

Table 29 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 29. Summary of the Overlap of Content Alignment between NAEP and ACCUPLACER Items and the NAEP Framework and ACCUPLACER Framework at the Standard Level

NAEP Framework	**NAEP Items** (40 items)		**ACCUPLACER Items** (55 items weighted as 70)[20]	
	Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
	% of Total Hits		**% of Total Hits**	
1 - Locate/Recall: Locate or recall textually explicit information within and across texts. . .	14	23	24	22
2 - Integrate/Interpret: Make complex inferences within and across texts	76	66	76	72
3 - Critique/Evaluate: Consider text(s) critically	10	11	0	5
ACCUPLACER Framework	**NAEP Items** (131 items)[21]		**ACCUPLACER Items** (55 items weighted as 70)	
	Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
A. Identifying main ideas	7	7	14	16
B. Direct statements/secondary ideas	26	27	21	20
C. Inferences	49	46	30	27
D. Applications	18	20	11	13
E. Sentence relationships	0	0	24	24

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

NAEP items were found to assess all NAEP standards (1–3) and four of the five ACCUPLACER standards (A–D). ACCUPLACER E, "Sentence relationships," was not assessed on NAEP. The ACCUPLACER standard is narrowly focused on the relationship between two sentences; the NAEP standards and objectives describe skills and capacities applied to "texts." Comprehension

[20] Each form consists of 35 items—20 variable items and 15 common items—for a total of 70 items. The complete item set for both forms was analyzed. For efficiency, the 15 common items were coded just once by panelists (panelists saw all 55 unique items), and the codes for the common items were weighted as double to retain the balance of content and complexity of the two forms (the full 70 items across both forms).

[21] The percentages in this table indicate the distribution of total hits. It should be noted that, as shown in Table 20, 12% and 15% of the adjusted total hits for NAEP items were determined by panelists to be uncodable to any ACCUPLACER objective.

of the relationship between two sentences may be indirectly assessed by many NAEP items (as part of the student's understanding of the larger text), but no NAEP standard or objective describes skills specifically applied at the sentence level. Eight and 10 of the 131 NAEP items analyzed for alignment to the ACCUPLACER framework were determined to be uncodable by Panels 1 and 2, respectively. Adjusting percentages for all uncodable ratings, 12% and 15% of the NAEP items were determined by the majority of panelists not to align to any ACCUPLACER objective.

ACCUPLACER items were found to assess all of the ACCUPLACER standards (A–E) and NAEP Standards 1 and 2. They did not cover NAEP Standard 3, "Critique/Evaluate," in Panel 1, and had limited coverage in Panel 2.

With regard to alignment to the NAEP framework, both ACCUPLACER items and NAEP items had a majority (66% and 76%) of hits to Standard 2, "Integrate/Interpret." The remaining NAEP item alignments in the short form of 40 items were split between Standard 1 and Standard 3, although the 10% and 11% of NAEP hits (4.29 mean hits) to Standard 3 was below the standard WAT threshold value of six items. In addition, it is worth noting that the items coded to Standard 3 were almost exclusively constructed-response items, and these items were not weighted for their multiple score-point value. ACCUPLACER had little to no coverage of NAEP Standard 3.

In relation to the ACCUPLACER framework, both assessments had the most hits to ACCUPLACER C, "Inferences." However, the NAEP items had a higher percentage of hits to ACCUPLACER C than did the ACCUPLACER items, with 46% and 49% of NAEP items compared to 27% and 30% of ACCUPLACER items matched to the standard. As previously noted, ACCUPLACER C is a very broad objective, overlapping at a high level of generality with many of the NAEP objectives for Standard 2, "Integrate/Interpret," all of which require making "complex inferences within and across texts." In addition, panelists applied a decision rule allowing NAEP vocabulary items to be aligned to ACCUPLACER C, based on the interpretation that using context to determine the meaning of words requires making inferences about connections between details, information, or ideas (Decision Rule 2 for the ACCUPLACER framework).

The NAEP items also had a greater percentage of hits than the ACCUPLACER items (18% and 20% compared to 11% and 13%) to ACCUPLACER D, "Applications," which includes understanding "how the author uses language to achieve his/her purpose," a skill addressed by several NAEP Standard 2 objectives.

For ACCUPLACER items, the percentage of hits to ACCUPLACER A, "Identifying main ideas" (14% and 16%), was approximately double that of the percentage of hits from NAEP items to the same objective (7%). In addition, approximately one-quarter of the ACCUPLACER hits (24%) and none of the NAEP items were to ACCUPLACER E, "Sentence relationships."

Overall, the NAEP items covered all of the ACCUPLACER objectives except ACCUPLACER E, but with a stronger emphasis on ACCUPLACER C and D. The ACCUPLACER items covered all of the ACCUPLACER objectives, with B, C, and E receiving the most emphasis. The NAEP items that did not align to the ACCUPLACER framework called

for evaluation or theme (or other specifically literary features) of a literary work, skills not addressed by any ACCUPLACER objective.

Overlap in content alignment to the NAEP framework can also be examined at the more finely grained objective level. Table 30 shows the overlap of alignment of each assessment to the NAEP framework in terms of the percentages of total hits.

Table 30. Summary of the Overlap of Content Alignment between NAEP and ACCUPLACER Items and the NAEP Framework at the Objective Level

NAEP Framework			**NAEP Items** (40 items)		**ACCUPLACER Items** (55 items weighted as 70)[22]	
			Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
Standard	**Goal**	**Objective**	**% of Total Hits**		**% of Total Hits**	
1	1.1	1.1.a	11	14	20	15
	1.2	1.2.a	1	0	0	0
		1.2.b	0	1	0	0
		1.2.c	0	0	0	0
		1.2.d	0	0	0	0
		1.2.e	0	0	0	0
	1.3	1.3.a	1	6	2	5
		1.3.b	0	0	0	0
		1.3.c	0	1	2	2
		1.3.d	0	0	0	0
2	2.1	2.1.a	0	0	7	10
		2.1.b	4	1	25	20
		2.1.c	2	0	1	5
		2.1.d	13	9	3	5
		2.1.e	1	1	0	1
		2.1.f	2	1	1	2
	2.2	2.2.a	2	3	1	2
		2.2.b	3	2	0	0
		2.2.c	6	7	0	0
		2.2.d	2	9	0	0
		2.2.e	0	0	0	0
	2.3	2.3.a	3	2	7	10
		2.3.b	16	7	28	10
		2.3.c	1	2	0	2
		2.3.d	0	0	0	0
		2.3.e	0	1	0	0
	2.4	2.4.a	20	20	3	6

[22] Each form consists of 35 items—20 variable items and 15 common items—for a total of 70 items. The complete item set for both forms was analyzed. For efficiency, the 15 common items were coded just once by analysts (analysts saw all 55 unique items), and the codes for the common items were weighted as double to retain the balance of content and complexity of the two forms (the full 70 items across both forms).

NAEP Framework			NAEP Items (40 items)		ACCUPLACER Items (55 items weighted as 70)[22]	
			Panel 1	Panel 2	Panel 1	Panel 2
Standard	Goal	Objective	% of Total Hits		% of Total Hits	
3	3.1	3.1.a	4	2	0	0
		3.1.b	1	0	0	4
		3.1.c	0	0	0	0
	3.2	3.2.a	0	0	0	0
		3.2.b	0	2	0	0
		3.2.c	0	0	0	0
	3.3	3.3.a	0	0	0	0
		3.3.b	5	4	0	0
		3.3.c	0	1	0	0
		3.3.d	0	0	0	0

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

As shown in Table 30, within the strong standard-level overlap at Standard 2, there is some variation in which objectives are assessed on each test. This table also illustrates how alignment at the objective level contributes to the range and balance described later in this section.

Categorical Concurrence

For alignment to the NAEP framework, the NAEP items were found to meet categorical concurrence for Standards 1 and 2. Categorical concurrence was not met for Standard 3, although the standard received 4.29 hits (10% and 11%). The ACCUPLACER items met categorical concurrence for Standards 1 and 2, but with only 3.7 mean hits (5%) from one panel only, the ACCUPLACER items did not meet categorical concurrence for Standard 3. In sum, both tests met the criteria for Standards 1 and 2; neither assessment met the threshold for categorical concurrence for Standard 3.

For alignment to the ACCUPLACER framework, the ACCUPLACER items were found to meet categorical concurrence for all objectives (A–E). The NAEP items met categorical concurrence for all objectives to which they aligned (A–D), but not for ACCUPLACER E, which received no hits.

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).

Depth-of-Knowledge Consistency

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 that were mapped to an objective in that standard were at or above the DOK level assigned to that objective. The ACCUPLACER items met depth-of-knowledge consistency only for Standard 1. Both panels found that the majority of ACCUPLACER items aligned to Standard 2 had a lower DOK level than that of the standard. Thus, NAEP had a higher DOK than ACCUPLACER for those items aligned to Standard 2.

For alignment to the ACCUPLACER framework, the ACCUPLACER items were found to meet depth-of-knowledge consistency in all objectives. The NAEP items met depth-of-knowledge consistency in the four objectives to which there were alignments, ACCUPLACER A–D, but not for ACCUPLACER E, to which no items were aligned.

Range of Knowledge

For alignment to the NAEP framework, for both NAEP and ACCUPLACER, only Standard 2 had a range of knowledge, with 50% or greater of the 17 objectives within that standard receiving alignments. For Standard 1, the NAEP items had alignments to 21% and 26% of the objectives, while ACCUPLACER items had alignments to 27% and 31% of the objectives in Standard 1. For Standard 3, NAEP items had alignments to 26% and 29% of the objectives, while ACCUPLACER items had alignments to 0% and 11% of the objectives in that standard.

For alignment to the ACCUPLACER framework, the ACCUPLACER items had hits to all five standards and NAEP items had hits to four of five standards. However, because all of the ACCUPLACER objectives exist at the same hierarchical level, the range of knowledge analyses are not applicable for this framework.

Balance of Representation

NAEP items had a balance of representation for all three standards in the NAEP framework. The ACCUPLACER items had a balance of representation for each of the two NAEP standards to which they were aligned. All of the ACCUPLACER objectives exist at the same hierarchical level, so the balance of representation analyses are not applicable for this framework.

Overall Conclusions

The following conclusions regarding the alignment of the 2009 NAEP Grade 12 Reading and the ACCUPLACER Reading Comprehension test can be drawn from the results of this alignment study.

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

The greatest commonality between the two tests is in their shared emphasis on the broad skills of comprehending and interpreting informational text, primarily through inferential reasoning. This is evident in the majority of items on both tests (66% and 76% for NAEP, and 72% and 76% for ACCUPLACER) matched to the NAEP standard “Integrate/Interpret: Make complex inferences within and across texts.” On both tests, the majority of alignments to “Integrate/Interpret” were to objectives that apply to informational text only or across both informational and literary texts.

The shared emphasis on the comprehension and interpretation of informational text can also be seen in the alignments on both tests to the ACCUPLACER framework. Although the ACCUPLACER standards do not explicitly refer to text type, they focus almost exclusively on elements typical of informational text. A majority of both NAEP and ACCUPLACER items were matched to ACCUPLACER “Inferences,” and both tests had significant percentages of

alignments to “Direct statements and secondary ideas” and “Applications.” A smaller percentage of items on both tests were aligned to “Identifying main ideas.”

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

As previously discussed, the alignments both within and across frameworks show that both tests emphasize the comprehension and interpretation of informational text, particularly through the use of inference. Within this broad area of convergence, however, there are differences in emphasis revealed in the alignments to specific objectives within both frameworks. In relation to the NAEP framework, the NAEP short-version items showed a far greater emphasis on the comprehension of vocabulary in context (Objective 4.a) and on the analysis of an author’s use of language (Objective 1.d). In relation to the ACCUPLACER framework, NAEP items showed more emphasis on the use of inference to interpret text (ACCUPLACER “Inferences”). The higher percentage of NAEP items aligned to “Applications” also reflects the greater emphasis in NAEP on understanding authors’ use of language.

In relation to the ACCUPLACER framework, the ACCUPLACER items showed a greater emphasis than the NAEP items on the identification of main ideas. In relation to the NAEP framework, the ACCUPLACER items showed more emphasis on the recall of specific details, facts, and information (NAEP 1.1.a).

In general, in the cross-framework alignments, the matches found in each test to the other’s framework (NAEP to ACCUPLACER and ACCUPLACER to NAEP) tended to be for the most general objectives within that framework. For example, the great majority of hits for ACCUPLACER items to NAEP objectives for “Integrate/Interpret” were to two of the most broadly stated NAEP objectives, “Draw conclusions” (2.3.b) and “Compare or connect ideas” (2.1.b). Many of the more specific NAEP objectives for “Integrate/Interpret,” such as “Find evidence in support of an argument” (2.2.c), received far fewer or no hits from ACCUPLACER items. Compared to ACCUPLACER, the NAEP items were more evenly distributed among NAEP objectives.

The majority of hits for NAEP items to ACCUPLACER standards were also to the broadest of those standards— “Inferences” and “Applications,” both of which overlap in content with a number of NAEP objectives but at a higher level of generality. The more specific ACCUPLACER standard, “Identifying main ideas,” received far fewer alignments from NAEP items.

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 reading subdomains are missing or not aligned?

In regard to differences in content, NAEP addresses reading skills related to both literary and informational text, while ACCUPLACER does not address reading skills specific to literary text. As expected, based on the framework-to-specifications comparison in the Interim Report, ACCUPLACER items had minimal matches to NAEP objectives for literary text. The main area

of alignment of ACCUPLACER items to the NAEP framework, NAEP objectives in Standards 1 and 2, applied to informational text only or to both informational and literary text.

The ACCUPLACER items also had minimal to no coverage of the NAEP standard "Critique/Evaluate." These findings are also consistent with the comparison of the two frameworks in the Interim Report; overall, the language of the ACCUPLACER objectives ("understand," "comprehend," "recognize") places more emphasis on comprehension and interpretation of text ("distinguish the main idea from supporting ideas" or "perceive connections between ideas made—implicitly—in the passage") than on critical analysis or evaluation ("Evaluate the strength and quality of evidence used by the author to support his or her position" in NAEP Objective 3.3.b, or "Judge the author's craft and technique" in NAEP Objective 3.1.a).

In regard to complexity, both assessments were found to meet the criteria for depth of knowledge consistency in relation to their own framework. In relation to the NAEP framework, however, only the NAEP items met the criteria for DOK consistency for all NAEP standards. The ACCUPLACER items met the criteria for depth of knowledge consistency only for NAEP "Locate/Recall." Although the majority of the ACCUPLACER item alignments were to objectives for NAEP "Integrate/Interpret," over half of these items were found to have a DOK level below that of the standard. In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the more in-depth reasoning required by DOK level 3. NAEP, by contrast, includes much longer reading passages and both multiple-choice and constructed-response items.

The NAEP covers skills specific to the comprehension and analysis of literary text while ACCUPLACER does not. In addition, NAEP covers the skills of evaluating and critiquing text, skills not addressed by ACCUPLACER. Finally, NAEP has a wider range of cognitive complexity than ACCUPLACER, with a substantially higher percentage of items at DOK level 3, requiring more in-depth analysis or evaluation. However, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

Overall, the NAEP items covered a broader range of cognitive complexity than the ACCUPLACER items. This is also apparent in the frameworks. The three NAEP standards, defined in terms of three different "cognitive targets" ("Locate/Recall," "Integrate/Interpret," and "Critique/Evaluate"), cover a broader range of cognitive complexity supported by the use of longer reading passages and the inclusion of both short and extended constructed-response items. The language of the ACCUPLACER standards ("understand," "comprehend," "recognize") places more emphasis on comprehension and interpretation of text (e.g., "distinguish the main idea from supporting ideas" in ACCUPLACER A, "Identifying main ideas," or "perceive connections between ideas made—implicitly—in the passage" in ACCUPLACER C, "Inferences") than on critical analysis or evaluation (e.g., "Evaluate the strength and quality of evidence" in NAEP 3.3.b, or "Judge the author's craft" in NAEP 3.1.a). In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the cognitive complexity typical of DOK Level 3 items. Although the NAEP items show a greater range of cognitive complexity and a greater emphasis on critical thinking, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

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 reading 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 to acquire the most complete, detailed framework available, and then to select the most appropriate grain size for coding and analysis, as was done in this study.

In this NAEP–ACCUPLACER reading 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 had only five broad categories (objectives) to code to in the ACCUPLACER framework, each accompanied by a brief description elaborating upon the intent of the objective. In contrast, panelists had 37 objectives to code to in the NAEP framework. As a result, whereas panelists could use the more specific NAEP objectives to inform their interpretations of the intent of that framework, they had less information with which to make judgments about alignment to ACCUPLACER. On the other hand, however, the larger number of NAEP objectives increased the likelihood that some objectives would not be matched to items.

As discussed in Sections III and V of this report, alignment across the NAEP and ACCUPLACER frameworks tended to occur among the broader objectives in each framework. Panelists were instructed to look for the best match to an objective, using the language of the objectives as their guide. Therefore, when the language of an objective was more specific, it might have been less likely for an item developed to another framework to precisely assess the described skill, and more likely for the item to assess a broader objective that encompasses the assessed skill. Thus, in order to create the most complete description of the alignment of the two assessments, every effort should be made to acquire the most complete, detailed framework available, and then to select the most appropriate level for coding and analysis.

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 reading. In particular, there was some discussion among panelists and facilitators about whether the simplest inferences in reading, such as those required by the use of synonyms, should be considered as DOK Level 1 rather than 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 frameworks 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, 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 that ran concurrent with this study. 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 WestEd's comprehensive reports on the NAEP–ACCUPLACER and NAEP–SAT mathematics studies, this presented a challenge to keeping to the allotted schedule. 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. (2010a). *ACCUPLACER®. Introduction for students.* Retrieved March 29, 2010, from http://www.collegeboard.com/student/testing/accuplacer/?print=true

The College Board. (2010b). *ACCUPLACER tests*. Retrieved September 8, 2010, from http://www.collegeboard.com/student/testing/accuplacer/accuplacer-tests.html

National Assessment Governing Board. (2008). *Reading framework for the 2009 National Assessment of Educational Progress.* Developed for the National Assessment Governing Board in support of Contract No. ED-02-R-0007, U.S. Department of Education, by American Institutes for Research.

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 Assessment Governing Board. (2009d). *Reading assessment and item specifications for the 2009 National Assessment of Educational Progress.* Prepared for the National Assessment Governing Board in support of Contract No. ED-02-R-0007, U.S. Department of Education, by American Institutes for Research.

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 ACCUPLACER mathematics core tests* (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 mathematics and SAT mathematics* (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).