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 SAT Critical Reading

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 SAT Critical Reading

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—one of four comprehensive reports to be submitted to the Governing Board—evaluated the alignment between the NAEP and SAT 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 SAT reading specifications, conducted in early 2010, and a series of alignment activities designed to investigate the degree of alignment between the pairs of assessments and frameworks/specifications.

These alignment activities were performed over the course of an alignment workshop conducted the week of March 8–12, 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 SAT assessment and the SAT reading framework, 3) the grade 12 NAEP and the SAT reading framework, and 4) the SAT 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. A short-version representative sample of items was used for the within-framework analyses (i.e., NAEP items to NAEP framework and SAT items to SAT framework). The complete NAEP item pool and the complete SAT item pool were used for the cross-framework analyses (NAEP items to SAT framework and SAT items to NAEP framework, respectively). 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 SAT?
- To what extent is the emphasis of reading content on NAEP proportionally equal to that on SAT?

- Are there systematic differences in content and complexity between the NAEP and SAT assessments in their alignment to the NAEP framework and between the NAEP and SAT assessments in their alignment to the SAT 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. The standards in each framework are listed below.

NAEP Framework Standards

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

SAT Framework Standards

A. "Sentence Completion"
B. "Passage-Based Reading"

NAEP Assessment to NAEP Framework Alignment

All NAEP items were determined to be codable to the NAEP framework. The NAEP short-version items (40 items) were found to assess all of the NAEP standards. NAEP items aligned primarily to "Integrate/Interpret," with the remaining NAEP item alignments divided between "Locate/Recall" and "Critique/Evaluate."

SAT Assessment to NAEP Framework Alignment

All of the 67 items from each of the two SAT forms were determined to be codable to the NAEP framework. SAT items were found to assess all of the NAEP standards, although the majority of SAT items were found to assess "Integrate/Interpret." The NAEP standards "Locate/Recall" and "Critique/Evaluate" each received minimal coverage.

SAT Assessment to SAT Framework Alignment

The SAT short-version items (46 items) were determined to be codable to the SAT framework. Approximately three-quarters of the SAT (short-version) alignments were to "Passage-Based Reading," while approximately one-quarter of the SAT total alignments were to "Sentence Completion."

NAEP Assessment to SAT Framework Alignment

All of the 131 items from the complete NAEP item pool were determined to be codable to the SAT framework. All NAEP item alignments were to "Passage-Based Reading," with no alignments to "Sentence Completion"; however, it is important to note that NAEP does assess vocabulary knowledge, a primary focus of the SAT "Sentence Completion" standard.

Categorical Concurrence

For alignment to the NAEP framework, the NAEP short-version items were found to meet categorical concurrence for all three standards. The SAT items met categorical concurrence for "Integrate/Interpret," but not for "Locate/Recall" or "Critique/Evaluate."

For alignment to the SAT specifications, the SAT short-version items were found to meet categorical concurrence for both standards. The NAEP items met categorical concurrence for "Passage-Based Reading," but not for "Sentence Completion."

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 aligned to an objective in that standard were at or above the DOK level assigned to that objective. The SAT items also met depth-of-knowledge consistency for all three NAEP standards.

For alignment to the SAT specifications, the SAT items were found to meet depth-of-knowledge consistency in both standards. The NAEP items met depth-of-knowledge consistency for "Passage-Based Reading," the only SAT standard to which they were aligned.

Range-of-Knowledge Correspondence

Range of knowledge correspondence is met for a standard if 50% or more of the objectives in that standard have items aligned to them. For alignment to the NAEP framework, the NAEP items met the typical WAT range of knowledge threshold criterion for "Integrate/Interpret." One panel found that "Locate/Recall" had a weak range of knowledge, while the other panel found that it did not meet the criterion for range; neither panel found "Critique/Evaluate" to meet the criterion. For the SAT items, range of knowledge was met for "Integrate/Interpret" but not for "Locate/Recall" or "Critique/Evaluate."

For alignment to the SAT specifications, the range of knowledge threshold criterion for "Passage-Based Reading" was met for both assessments. Regarding "Sentence Completion," it is important to note that, with only one objective in this standard, range of knowledge can be met with one mean item alignment. Therefore, range of knowledge is not applicable when referring to this standard.

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.

The NAEP short-version items met or weakly met the balance of representation criteria for all three standards in the NAEP framework. SAT items had very few alignments to the NAEP standards "Locate/Recall" and "Critique/Evaluate," but both panels found that balance of representation was met on both SAT forms for "Integrate/Interpret."

For alignment to the SAT framework, the SAT short-version items met the balance of representation criteria for "Passage-Based Reading," and the NAEP items weakly met the criteria for this standard. Regarding the SAT "Sentence Completion" standard, it is important to remember that, with only one objective in this standard, balance of representation can be met with one mean item alignment. Therefore, balance of representation is not applicable when referring to this standard.

Overall Conclusions

The following conclusions regarding the alignment of the 2009 NAEP Grade 12 Reading and the SAT Critical Reading 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 SAT?

The greatest commonality between the two tests is their shared emphasis on the broad skills of integrating and interpreting both informational and literary texts. This is evident in the majority of items from both tests aligned to NAEP Standard 2, "Integrate/Interpret," including many to Goal 2.1, "Make complex inferences within and across *both literary and informational texts*."

Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), there is also considerable overlap at the level of more specific skills.

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

Both tests had many of their item alignments to the same NAEP "Integrate/Interpret" objectives, often with similar percentages of alignments. Although there were some differences in emphasis, both tests also had notable percentages of alignments to SAT Objectives B.1.1–B.1.3 and B.1.5. Skills with overlap include inferring/analyzing the following:

- the "main idea" and "author's purpose" (SAT Objective B.1.1 and NAEP Objectives 2.3.a and 2.1.f);
- the "tone and attitude" of an author or character (NAEP Objectives 2.2.a and 2.2.c and SAT Objective B.1.4);
- the use of "rhetorical strategies" (NAEP Objective 2.1.d and SAT Objective B.1.2); and
- connections between ideas, perspectives, or problems (NAEP Objective 2.1.b and SAT Objectives B.1.3 and B.1.5).

Additionally, in the area of greatest content overlap—items on both tests aligned to objectives for NAEP "Integrate/Interpret" and aligned to SAT "Passage-Based Reading" Objectives B.1.1–B.1.5—both tests met the typical threshold criteria for depth of knowledge consistency; that is,

most of the items were coded at or above the DOK level of the objectives to which they were aligned.

Despite these similarities, there are some notable differences in emphasis between the two assessments. Both tests assess vocabulary skills. However, NAEP addresses vocabulary exclusively in the context of passage comprehension, while the majority of SAT vocabulary items are in a sentence-completion format, in which context plays a more limited role. This difference reflects NAEP's emphasis on the understanding of word meaning in context; the assessment is not intended to measure students' prior knowledge of word definitions. The SAT sentence-completion items provide some context within the single sentence text, but in many cases, students' success on the items almost certainly depends on their prior knowledge of word definitions.

In addition, panelists found considerably less emphasis in SAT than in NAEP on literal comprehension and critical evaluation, particularly the evaluation of the quality or effectiveness of an author's writing, skills covered in the NAEP standards "Locate/Recall" (locating/recalling specific details and features of texts) and "Critique/Evaluate" (evaluating texts from a critical perspective), respectively. This difference suggests a greater emphasis on these skills in NAEP.

Even with the minimal coverage of NAEP "Locate/Recall" and "Critique/Evaluate" standards by SAT items, all NAEP items found a match in the SAT framework. However, the broad language of the SAT framework can encompass the range of the NAEP items. For example, SAT Goal B.2, "Literal Comprehension," refers to items that "ask what is being said" in a "small but significant portion of a reading passage," a description that can easily accommodate most NAEP "Locate/Recall" items and objectives. In fact, nearly all items on the NAEP short version that were coded to "Locate/Recall" objectives in the NAEP framework were matched to SAT Goal B.2 in the SAT framework.

Similarly, SAT Objective B.1.3, to which approximately one-quarter of NAEP items aligned, includes "Evaluation," the primary focus of NAEP "Critique/Evaluate." The description in SAT Objective B.1.3 of items that "ask the test taker to evaluate ideas or assumptions in a passage" is compatible at a very general level with NAEP "Critique/Evaluate" objectives addressing the author's point of view, logic, or use of evidence. SAT Objective B.1.2, "Rhetorical Strategies," is also broad enough in its language to make it a reasonable match for some NAEP "Critique/Evaluate" items focused on "author's craft" or use of "literary devices." In the NAEP short version, all items that aligned to "Critique/Evaluate" objectives in the NAEP framework were aligned to either SAT Objectives B.1.2 or B.1.3, or both.

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

With regard to differences in content as described in the NAEP framework, SAT items had limited coverage of the knowledge and skills described by the NAEP standards "Locate/Recall" and "Critique/Evaluate." This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In

comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.

With regard to differences in complexity, NAEP items and objectives had a range of depth of knowledge including items at DOK Levels 1, 2, and 3, while SAT items and objectives were coded primarily at Levels 2 and 3.

Overall, the alignment results across the two sets of items and frameworks show a strong area of overlap in their coverage of SAT "Passage-Based Reading" objectives and NAEP "Integrate/Interpret" objectives, as well as some important differences.

I. Introduction

Purpose

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

The Technical Panel's recommended multi-method approach (National Assessment Governing Board, 2009c) includes conducting content alignment studies 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).[1] 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 SAT Critical Reading test (SAT). Alignment findings from the studies of the alignment between NAEP and SAT Mathematics, ACCUPLACER Reading Comprehension, and ACCUPLACER Mathematics Core Tests are presented in separate reports (WestEd, 2010a, 2010b, 2010c).

[1] Preliminary comparability studies were conducted by the Educational Testing Service for use by the National Assessment Governing Board and the College Board to determine the feasibility of relating NAEP and SAT in mathematics and reading and of examining alignment of the two more fully (Pitoniak, Reese, & Tannenbaum, 2008a, 2008b).

Governing Board's Approach to Preparedness

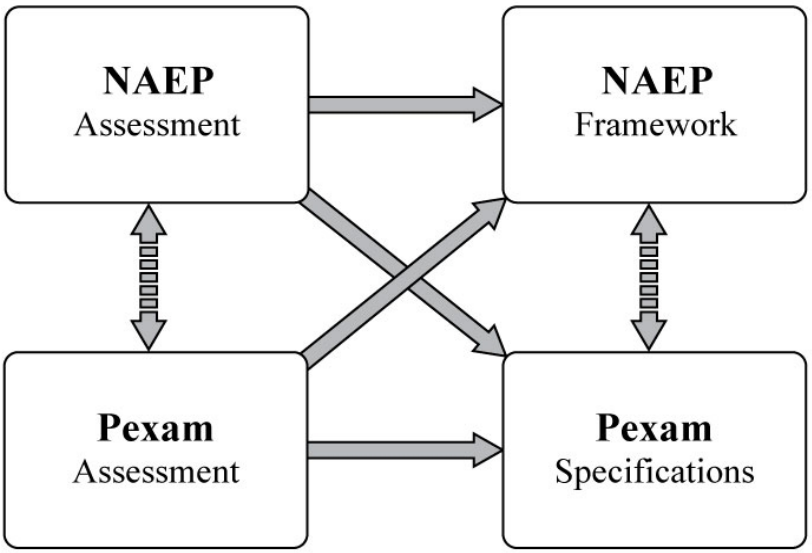
The Governing Board is focusing its conceptualization of 12th grade preparedness on academic qualifications and does not propose to address a range of behavioral and attitudinal aspects of student performance in postsecondary activities that are not measured by NAEP (e.g., time management skills, diligence). The Governing Board further limits its definition of postsecondary preparedness to refer to the academic skills required for placement into entry-level college-level credit courses that count toward a four-year undergraduate degree, or for placement into military or civilian job training programs[2] (e.g., apprenticeship programs, vocational institute or certification programs, on-the-job training programs), with no prediction of success in such college-level courses or job training programs.

Assessment-to-Assessment Alignment

While a typical alignment study explores the alignment between an assessment and a set of standards, the Technical Panel called for studies that would investigate the degree to which NAEP is aligned in content and complexity to other assessments, assessments that were developed from different frameworks for different purposes. To accomplish this objective, the Governing Board contracted with Dr. Norman Webb to propose a bi-directional, multifaceted study design to look at alignment between an assessment and its own framework (e.g., NAEP with NAEP) and between an assessment and another assessment's framework or set of specifications (e.g., NAEP with SAT), as illustrated in Figure 1 on the following page. (The full text of the resulting study design document is provided in Appendix A.) This study design comprises both a qualitative comparison of the NAEP reading framework and the SAT Critical 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).

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

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



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

Although both NAEP and SAT 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. SAT tests students' knowledge of reading, writing, and mathematics at the high school level and is primarily used by colleges and universities to help determine individual students' academic readiness for college (College Board, 2010).

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.

[3] In the design document, the term "Pexam" is the generic term used for the performance exams to which NAEP is compared in the series of alignment studies.

This report addresses the following specific questions:

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

Alignment Study

The NAEP–SAT 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 SAT Critical Reading specifications occurred in early 2010, while the alignment activities were performed over the course of an alignment workshop conducted the week of March 8–12, 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 SAT Critical Reading assessment and the SAT Critical Reading specifications, 3) the grade 12 NAEP and the SAT Critical Reading specifications, and 4) the SAT Critical Reading 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, this NAEP–SAT reading alignment study was conducted concurrently with the NAEP–SAT 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.[4] The development of the SAT Critical 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 SAT 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 SAT 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;
- 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.

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

II. Methodology

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

Study Design Overview

This 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. Norman Webb for the Governing Board (National Assessment Governing Board, 2009b) to guide grade 12 NAEP alignment studies in evaluating the degree to which the grade 12 NAEP reading assessment aligns in content and complexity to the SAT Critical Reading assessment.

The study design called for a qualitative comparative analysis of the similarities and differences between the NAEP and SAT frameworks. The result of this analysis is the NAEP–SAT 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 SAT Critical Reading assessment and the SAT framework, 3) the grade 12 NAEP and the SAT framework, and 4) the SAT Critical Reading 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). The within-framework baseline alignment was important in interpreting the degree of cross-framework alignment.

The alignment methodology employed in this study called for each 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 following.

1. NAEP framework reviewed and coded for DOK
2. NAEP items coded to NAEP framework

3. SAT framework reviewed and coded for DOK
4. SAT items coded to SAT framework
5. NAEP items coded to SAT framework
6. SAT items coded to NAEP framework

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

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 pilot 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, further 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, and is included in Appendix C.

The SAT reading framework represents the assessed content as two broad content categories: sentence completion and passage-based reading. The passage-based reading category is further defined by three categories: extended reasoning (comprising five sub-categories), literal comprehension, and vocabulary in context. 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 detailed content information intended to elucidate the intent of each category. Additionally, WestEd added alphanumeric coding to the content categories corresponding to standard (e.g., B), goal (e.g., B.1), and objective (e.g., B.1.1) levels. It should be noted, however, that what are referred to in this report at the highest level as SAT "standards" are categories based on item types, rather than standards based on cognitive targets, as the NAEP "standards" used in this study are. Panelists were instructed to code to the most specific level possible (the objective level); due to the organization described above, depending on the content, the most specific level could be one (i.e., A), two (e.g., B.2), or three (e.g., B.1.1) levels into the framework. The SAT framework used in this study is included in Appendix C.

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

Comparison of Critical Features of the Assessments

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

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

	NAEP Grade 12 Reading Assessment	SAT Critical Reading 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 are intended to assess all three cognitive targets.
		SAT Critical Reading section is organized by two broad content categories with subcategories as follows:
		• Sentence Completion (28%)
		• Passage-Based Reading (72%)
		  • Extended reasoning
		  ▪ Primary purpose
		  ▪ Rhetorical strategies
		  ▪ Implication and evaluation
		  ▪ Tone and attitude
		  ▪ Application and analogy
		  • Literal comprehension
		  • Vocabulary in context
		Questions assess students' reading skills, such as:
		• Identifying main and supporting ideas
		• Determining the meaning of words in context
		• Understanding authors' purposes
		• Understanding the structure and function of sentences

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	The passage-based reading questions are based on 7 passages: • one paired paragraphs • two paragraph reading • one 500-word passage* • one 650-word passage* • two 800-word passages* *Note: One of the long passages is a pair of related passages (i.e., instead of an 800-word passage, there will be two related 400-word passages, etc.). SAT reading passages are taken from the following fields: • Natural sciences

	NAEP Grade 12 Reading Assessment	SAT Critical Reading Assessment
	Informational texts (70%) • *30% Exposition:* e.g., essay, literary 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	• Humanities • Social sciences • Literary fiction
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 from 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 become more	SAT reading passages: • Have narrative, argumentative, or expository elements • May be paired with related passages on a shared theme or issue • Often have line numbers or numbered elements that are then referenced in the questions that follow

	NAEP Grade 12 Reading Assessment	SAT Critical Reading Assessment
	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 range in length from 100 to about 850 words.
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 the reading difficulty of SAT 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	Passage-based reading questions will ask students to determine the meanings of words from their context. Sentence completion questions will measure students' knowledge of the meaning of words and understanding of how the different parts of a sentence fit together logically: • Each question presents students with a sentence that has one or more blanks. • Students must choose the word(s) to complete the sentence that best fit the meaning of the sentence as a whole.

	NAEP Grade 12 Reading Assessment	SAT Critical Reading 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	Each form consists of 67 items: • Sentence Completion: 19 items • Passage-Based Reading: 48 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.
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	No publicly available information, and none of the information furnished for this study describes the time per item type.
Assessment Time	Each student spends approximately 50 minutes (two blocks at 25 minutes each) taking the NAEP Reading Assessment.	Each student has 70 minutes (two 25-minute sections and one 20-minute section) to take the SAT Critical Reading Assessment.
When Given	NAEP assesses and reports grade 12 reading results every four years.	SAT is offered seven times a year in the U.S. and six times at international sites.
Testing Population	The 2009 Grade 12 NAEP was administered to: • 48,900 12^{th} grade students in reading in 1500 public schools • Random samples of students designed to be representative of the nation	SAT is administered to high school students planning to attend college or university.

	NAEP Grade 12 Reading Assessment	SAT Critical Reading Assessment
	• 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	The College Board's Services for Students with Disabilities (SSD) provides a range of accommodations, such as: • Braille tests • Large print • Extra/extended breaks • Sign language interpreters • Extended time For a complete list of SAT reading accommodations, see: http://www.collegeboard.com

Item Scoring	The items are scored as:
	• Multiple choice:
	  • Incorrect 0
	  • Correct 1
	• Short constructed response:
	  • Incorrect 0
		The items are scored as:
		• Multiple choice:
		  • Incorrect: ¼ point is subtracted
		  • Correct: 1 point is added
		No points are subtracted for omitted questions.

	NAEP Grade 12 Reading Assessment	SAT Critical Reading Assessment
	• 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 reported 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 such knowledge to real-world situations, and analytical skills appropriate to the	**Scaled scores:** Range of 200–800 A raw score is calculated: • Total points answered wrong are subtracted from the number answered correctly. • If the resulting score is a fraction, it is rounded to the nearest whole number. Raw score is converted to the 200–800 scaled score by a statistical process called equating: • Adjusts for slight differences in difficulty between test editions • Ensures that a student's score does not depend on how well others did on the same edition of the test

	NAEP Grade 12 Reading Assessment	SAT Critical Reading Assessment
	subject matter. • **Advanced —** This level signifies superior performance. Test scores and achievement levels are used to report on the performance of grade 12 students nationally. In 2009, 11 states participated in the first pilot for reporting state NAEP results at grade 12.	

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[5]—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).

The SAT Critical Reading assessment is a fixed-form test comprising 67 items. After collaboration with WestEd and the Governing Board to determine the optimal item pool to use in this study, the College Board provided two disclosed forms (Forms A and F) consisting of 134 unique multiple-choice items (67 items per form).

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-

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

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 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.[6] The Governing Board reviewed and approved this short-version NAEP item pool for use in aligning to the NAEP framework.

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

SAT-to-SAT Alignment
To reduce coding time given scheduling constraints, a subset ("short version") consisting of 46 items was analyzed for alignment to the SAT framework. Following review of the SAT forms and using the item-level characteristics provided by the College Board, WestEd selected these 46 items to be representative of the range of items in the two forms. The resulting sample item pool was a reasonable approximation of a representative sample, balancing the number of items with the following characteristics:

- item type (sentence completion or passage-based content);
- complexity;
- content classification; and
- passage length.

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

Additionally, all items associated with each selected passage were used. Since the item booklet for Form A was used in both the SAT–SAT and SAT–NAEP studies, the items in Form A were reordered and numbered sequentially so that the 46 items selected for the short-form sample appeared first, followed by the remaining 21 items. The items in Form F were numbered in their original order and contained in a separate booklet used in the SAT–NAEP study.

SAT-to-NAEP Alignment
All 134 items from the two SAT forms (A and F) were analyzed for alignment to the NAEP framework. The two forms were analyzed separately to preserve the proportional distribution of content on the two fixed forms.

For alignment purposes, within the WAT system, NAEP items were numbered sequentially, with the 40 items from the short-form sample appearing first. The two SAT forms were set up as separate studies in the WAT, each with its own sequential numbering.

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 SAT), and, in so doing, evaluate the degree of alignment of two assessments by comparing how the items on the two assessments represent their respective content domains.

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

Alignment Criteria Used in the Study

The alignment methodology employed in this study used four criteria to determine the degree of alignment between the NAEP and SAT assessments and the NAEP and SAT 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 SAT 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 the NAEP and SAT assessments as well as the NAEP and SAT 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 SAT tasks.

Adjudication Discussions Implemented in the Study

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

Adjudication of DOK of Objectives

As directed by the study's design document (National Assessment Governing Board, 2009b, p. 13), both reading panels were required to reach joint agreement on the DOK levels of each assessment framework's objectives.[7] 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 the 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 to determine whether differences were based on a misinterpretation or systematic difference in application of the protocol, were related to specific issues with an item or standard, or were random differences among panelists.

Cross-Panel Adjudication
The facilitators then met separately with WestEd project staff and, usually, the Governing Board's Contracting Officer's Representative (COR), to compare the results of the two groups

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

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.

Alignment Procedure Implemented in the Study

This alignment workshop occurred over five consecutive days. A full agenda by day is provided in Appendix D, 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. A representative from the College Board was invited to present an overview of SAT but was unable to attend. A copy of the PowerPoint presentation shared during this introductory session can be found in Appendix E. Following this introductory session, panels from the two content areas separated; for the remainder of the week, they reconvened as a whole group only for daily announcements prior to the start of each day's alignment activities, if necessary.

Following the introductory session, the 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 F).

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

Review NAEP Framework and Assign DOK Levels to Each Objective

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

Map NAEP Items to the NAEP Framework

Prior to mapping NAEP items to the NAEP framework, the combined 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 SAT assessment items.[8] 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 panelists through the coding of a limited sample set of active NAEP items[9] 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 SAT framework later in the week, by keeping a hard-copy record of each item's DOK level.

[8] 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 released SAT items included in the *Skills Insight* document (The 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.

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

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. Panelists then entered changes to their codes, if any, into the WAT tool. 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 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 their judgment of the coding had changed. 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 the SAT 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 SAT 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 SAT items to the SAT framework) should occur before cross-framework coding (i.e., mapping NAEP items to the SAT framework, and mapping SAT 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 SAT framework 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 consensus on DOK levels; the facilitators met separately to identify and adjudicate differences between the two groups to achieve cross-panel consensus on DOK levels; and, upon reaching cross-panel consensus, the facilitators communicated these values to their panelists and entered SAT framework DOK values into the WAT.

Map SAT Items to the SAT Framework

As with the mapping of NAEP items to the NAEP framework, a subset of 46 SAT items was selected to be mapped to the SAT 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 SAT items[10] from the item booklet to ensure calibration among panelists. Once calibration was reached, panelists began to independently map the remaining SAT items to the SAT 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. 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 protocol or the framework or misinterpretations of the protocol, framework, or items.

Map NAEP Items to the SAT Framework

The procedures described earlier for mapping each assessment's items to its framework were used to map NAEP items to the SAT framework, although for this alignment task the entire NAEP item pool was used. Because the short set of 40 NAEP items had already 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 SAT 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.

[10] 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.

Map SAT Items to the NAEP Framework

The procedures described earlier were used to map SAT items to the NAEP framework objectives. Because the short set of 46 SAT items had been assigned DOK levels when being mapped to the SAT framework, those DOK levels were re-entered into the WAT for this task; thus, the task of mapping the first 46 SAT items to the NAEP framework was limited to determining alignment to objectives. For all remaining SAT items, within this task, DOK levels were assigned and alignment to objectives was determined.

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.

SAT Reading Framework for Alignment: Decision Rules

1. Items based on multiple passages will be interpreted as aligning to "Applications" when such items require students to use information or ideas from "passage A" in order to comprehend, interpret, analyze, or evaluate "passage B." Items which only ask students about similarities or differences between passages—without requiring them to apply information or ideas from one to the other—will not be considered to align to "Applications."
2. The term "rhetorical strategies" will be interpreted as including literary devices or techniques.

SAT Items to NAEP Reading Framework: Decision Rules

1. SAT sentence-completion items will be interpreted as aligned to NAEP Objective 2.4.a on the grounds that all such items require students to determine the meaning of words in context.

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 SAT 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 SAT were disqualified from participation in the alignment workshop. In addition, WestEd sought panelists who would represent a range of knowledge of each assessment on each panel.

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

- Referrals from the 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.[11]

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

- On the first panel, 43% of panelists (3 of 7) reported experience in both secondary and postsecondary reading education; 14% (1 of 7) had secondary teaching experience only; and 43% (3 of 7) had postsecondary teaching experience only.
- On the second panel, 57% of panelists (4 of 7) reported experience in both secondary and postsecondary reading education; 14% (1 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 SAT assessments. Every attempt was made to balance each panel by geographic representation, race, ethnicity, and gender, although panelist availability limited the results of these attempts. The distribution of gender was comparable across the panels, with six women on each of the seven-member panels, as was representation of advanced degrees, with five doctoral degrees represented on Panel 1 and four doctoral degrees represented on Panel 2. Panelists represented a range of geographic areas, including the Northeast (New York, New Jersey, Pennsylvania), the Southeast (Kentucky, North Carolina, Louisiana, Virginia), the Midwest (Indiana, Iowa, Michigan), the Southwest (New Mexico), the Northwest (Oregon), and the West (California). Regarding diversity of race/ethnicity, panelists identified themselves as White/Caucasian/of European descent (8), Black/African-American (2), Native American or Alaskan and White/Caucasian/of European descent (1), Native American or Alaskan and Hispanic or Latino (1), and Eastern European (1). One panelist did not identify her race/ethnicity. A list of panelists organized by panel follows.

Reading Panel 1

[redacted]

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

[redacted]

Reading Panel 2

[redacted]

[redacted]

[redacted]

Preparation, Materials, and Logistics

Facilitator Training

Prior to this alignment workshop, a 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–SAT interim report. The facilitators were also asked to review the NAEP and SAT 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 training:

- Alignment workshop objectives and study design overview
- Agenda review
- NAEP and SAT assessment overview and discussion of issues
- NAEP and SAT framework overview and discussion of issues
- Discussion of NAEP and SAT decision rules
- Panelist training
- Facilitator roles and responsibilities (e.g., security protocols)
- Cross-panel adjudication worksheet
- Study launch page and electronic surveys
- WAT system use

Materials from both facilitator training sessions, as well as a facilitator process reference sheet are included in Appendix G.

Pre-Workshop Facilitator and Panelist Materials

In preparation for the NAEP–SAT study, the study's lead facilitator developed the comparative analysis to document the similarities and differences between the NAEP framework and the SAT framework. Prior to this alignment workshop, the facilitators reviewed the 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 SAT 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), and College Board's *SAT® Skills Insight™* (College Board, 2008). In an accompanying cover letter, panelists were asked to review the documents prior to the start of the alignment workshop to ensure that they were familiar with the content of the assessments.

Facilitator and Panelist Binder Materials

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

Panelist Training Materials

Panelist training for assigning DOK levels to objectives occurred on the first morning of the alignment workshop. Panelist training for assigning DOK levels of items and for coding items to objectives occurred on the second morning of the workshop. In addition, facilitators reviewed the alignment criteria at the beginning of each alignment session and provided refresher training as needed. A combined (reading and mathematics) panel training session introduced the purpose of the overall study and the NAEP and SAT assessments; it also provided an overview of the alignment process, definitions of alignment criteria, and use of the WAT (copies of panelist training materials are provided in Appendix E). Following this introduction and overview, the two reading panels relocated to a separate room and 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 *SAT® Skills Insight™ Critical Reading Real SAT Questions and Answers* (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 E.

On-Site Security of Materials

WestEd secured framework, 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 SAT assessments. For NAEP, all 131 items were organized by passage block and numbered sequentially within each block, with the 40 items identified for coding to the NAEP framework included first. 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.

For SAT, the items from the two forms were bound separately. In the Form A booklet, the short version of 46 items was listed first, followed by the remaining 21 Form A items. The Form F booklet retained the original item sequence. 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.

WestEd staff made available individual copies of the NAEP and SAT 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 F. Process questionnaires are discussed in Section IV of this report.

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

Debrief Questionnaires

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

Process Questionnaires

- A training questionnaire was administered following panelists' training on the first day of the alignment workshop to solicit feedback on the training's effectiveness and to identify areas in which more information might be needed. This questionnaire was administered via the online SurveyMonkey system (SurveyMonkey).[12]
- An evaluation-of-process questionnaire was administered at or near the end of each of the second, third, and fourth days of the alignment workshop. These questionnaires were used to monitor panelists' understanding of the process and to solicit questions, concerns,

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

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

Final Debrief

As the final task of the week, the combined panels convened with the two facilitators, WestEd staff, and Governing Board observers to discuss how the process captured the content similarities/differences between the assessments, to what degree the two assessments aligned, and, considering the items in each assessment, how the 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 SAT items (i.e., assigned item numbers and item weights) and frameworks into the WAT and created the five requisite WAT studies for each group:

- NAEP (short version) items to NAEP framework
- SAT (short version) items to SAT framework
- NAEP items to SAT framework
- SAT (Form A) items to NAEP framework
- SAT (Form F) items to NAEP framework

During the workshop, WAT system server errors outside the control of the project were experienced for one of the reading WAT studies, making it impossible to download certain WAT reports for this study. A duplicate study was created in the WAT, and the raw data table was used to re-enter the data.

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–SAT 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.

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. Although some of the recommendations are specific to the ACCUPLACER assessment, all recommendations are included here to preserve the completeness of the list and because of the potential for lessons learned to be applied to the SAT study.

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.[13]

Facilitator Training

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

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

- 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.

Schedule

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

Equipment/Technology

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

Analysis

- Clarify and document the process for averaging or aggregating results across the two panels outside the WAT.
- Combine the ACCUPLACER forms into one item pool for the operational studies, including the common items only once, in their first position, and assign them a double

weighting to retain the accuracy of the proportions. Make cross-assessment comparisons at the item pool level.[14]

All recommendations were implemented.

[14] This recommendation is relevant to ACCUPLACER only. For SAT reading, the two forms were analyzed separately. Because there was no overlap of items across the core tests, no weighting was required.

III. Alignment Results

This section presents the results of the NAEP-to-SAT alignment study. The section begins by reporting the interrater agreement within panels. Then, the DOKs of the frameworks and 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). Values are considered in the highest range in which they fall; values greater than 0.7 reflect adequate agreement, while values greater than 0.8 reflect good agreement. Because low variance among the items can make the intraclass correlation statistic misleading, the WAT also provides pairwise comparison values (p. 115). The WAT calculates pairwise comparison for DOK by comparing the ratings assigned by each possible pair of panelists in a panel, dividing the number of agreeing pairs by the total number of pairs, and then finding the average agreement across all items on a test. Values of 0.7 or higher reflect good agreement, values of 0.6 or higher reflect reasonable agreement, and values lower than 0.5 reflect poor agreement (p. 116).

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

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

Table 2. Interrater Agreement of Panels by Sub-Study

Sub-Study	Panel 1	Panel 2
Sub-Study 1: NAEP to NAEP	*DOK* Intraclass Correlation: 0.942 Pairwise Comparison: 0.6643 *Objective, Standard* Objective Pairwise Comparison: 0.5863 Standard Pairwise Comparison: 0.9255	*DOK* Intraclass Correlation: 0.96 Pairwise Comparison: 0.7357 *Objective, Standard* Objective Pairwise Comparison: 0.5519 Standard Pairwise Comparison: 0.8794
Sub-Study 2: SAT to NAEP	*Form A* *DOK* Intraclass Correlation: 0.8833 Pairwise Comparison: 0.7157 *Objective, Standard* Objective Pairwise Comparison: 0.6659 Standard Pairwise Comparison: 0.9374 *Form F* *DOK* Intraclass Correlation: 0.8421 Pairwise Comparison: 0.6304 *Objective, Standard* Objective Pairwise Comparison: 0.5835 Standard Pairwise Comparison: 0.9238	*Form A* *DOK* Intraclass Correlation: 0.8996 Pairwise Comparison: 0.7825 *Objective, Standard* Objective Pairwise Comparison: 0.6167 Standard Pairwise Comparison: 0.8678 *Form F* *DOK* Intraclass Correlation: 0.8033 Pairwise Comparison: 0.6873 *Objective, Standard* Objective Pairwise Comparison: 0.6312 Standard Pairwise Comparison: 0.8974
Sub-Study 3: SAT to SAT	*DOK* Intraclass Correlation: 0.913 Pairwise Comparison: 0.7681 *Objective, Standard* Objective Pairwise Comparison: 0.6827 Standard Pairwise Comparison: 0.9689	*DOK* Intraclass Correlation: 0.8957 Pairwise Comparison: 0.7681 *Objective, Standard* Objective Pairwise Comparison: 0.674 Standard Pairwise Comparison: 0.9266
Sub-Study 4: NAEP to SAT	*DOK* Intraclass Correlation: 0.9283 Pairwise Comparison: 0.7019 *Objective, Standard* Objective Pairwise Comparison: 0.6347 Standard Pairwise Comparison: 0.9935	*DOK* Intraclass Correlation: 0.9544 Pairwise Comparison: 0.7579 *Objective, Standard* Objective Pairwise Comparison: 0.7264 Standard Pairwise Comparison: 1

Looking across panels, Table 2 shows that interrater agreement (within-panel) values for each panel appeared comparable. Interrater agreement for DOK (intraclass correlation and pairwise comparison) met the threshold value for "good" for all sub-studies for both panels, except for Sub-Study 1 in Panel 1, where pairwise comparison was "reasonable" as defined by the WAT. Likewise, standard pairwise comparison values met the "good" threshold for all sub-studies for both panels. For match to objective, objective pairwise comparison values met the "good" or "reasonable" threshold for all sub-studies, with three exceptions. For both panels in Sub-Study 1

(NAEP-to-NAEP) and for Panel 1's SAT Form F alignment to the NAEP framework in Sub-Study 2, objective pairwise comparison was just below the "reasonable" range as defined by the WAT, with 0.5863, 0.5519, and 0.5835, respectively. Lower objective pairwise comparison values can result from overlapping or unclear objectives, as well as from multiple coding of items. For all three of these sub-studies, agreement at the standard level was very high (0.9255, 0.8794, and 0.9238, 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, the degree of cross-panel agreement attained was monitored throughout the study, as stipulated in the study design document. Where specific points of discrepancy and adjudication occurred, these are discussed in the context of each sub-study.

DOK Levels of the NAEP and SAT Frameworks

Panelists assigned DOK levels to each objective in the NAEP and SAT frameworks. The within-panel DOK ratings were then compared across panels and the two facilitators reached consensus on the final DOK ratings for each objective, discussing with the combined reading panels as appropriate. Consensus DOK values for the NAEP and SAT frameworks are shown in Tables 3 and 4, respectively. DOK ratings were assigned to the 37 NAEP objectives and eight SAT objectives. These ratings are reported at the goal and standard level in the tables. DOK ratings for each objective can be found in Appendix C.

Table 3. DOK Findings for the NAEP Reading 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	-	2 (33%)	4 (67%)	2.67
2.2	5	-	-	5 (100%)	3
2.3	5	-	3 (60%)	2 (40%)	2.4
2.4	1	-	1 (100%)	-	2
2 overall	**17**	**-**	**6 (35%)**	**11 (65%)**	**2.65**
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%)**	**6 (16%)**	**21 (57%)**	**2.3**

As shown in Table 3, all objectives in NAEP Standard 1, "Locate/Recall," were assigned a DOK level of 1, for an average DOK level of 1. For Standard 2, "Integrate/Interpret," objectives were assigned DOK levels of 2 or 3, for an average DOK level of 2.65. For Standard 3, "Critique/Evaluate," 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 2.3. Across all standards and the 37 objectives, the distribution of DOK levels was 27% (10) at Level 1, 16% (6) at Level 2, and 57% (21) at Level 3. The standard with the highest percentage of Level 3 items and the highest average overall DOK was "Critique/Evaluate."

Table 4. DOK Findings for the SAT Reading Framework

SAT Framework	# of Objectives	# and % of Obj. at DOK 1	# and % of Obj. at DOK 2	# and % of Obj. at DOK 3	Average DOK
A	1	-	1 (100%)	-	2
A overall	**1**	**-**	**1 (100%)**	**-**	**2**
B.1	5	-	1 (20%)	4 (80%)	2.8
B.2	1	-	1 (100%)	-	2
B.3	1	-	1 (100%)	-	2
B overall	**7**	**-**	**3 (43%)**	**4 (57%)**	**2.57**
ALL	**8**	**-**	**4 (50%)**	**4 (50%)**	**2.5**

As shown in Table 4, SAT A, "Sentence Completion," was assigned a DOK level of 2. For SAT B, "Passage-Based Reading," objectives were assigned DOK levels of 2 (43%) or 3 (57%), for an average DOK level of 2.57. Overall, the average DOK level of all SAT objectives was 2.5. Across all standards and the eight objectives, the distribution of DOK levels was 50% (4) at Level 2 and 50% (4) at Level 3. The standard with the highest percentage of Level 3 objectives and the highest average overall DOK was "Passage-Based Reading." All four "Passage-Based Reading" objectives coded at Level 3 were within Goal B.1, "Extended Reasoning."

Comparing Tables 3 and 4, the objectives in the NAEP framework were found to have an average DOK level of 2.3, compared with an average of 2.5 in the SAT framework. In terms of emphasis of DOK, NAEP has 27% of objectives at Level 1 and SAT has no objectives at Level 1. NAEP has 16% of objectives at Level 2, compared with 50% of SAT objectives at Level 2. NAEP has 57% of objectives at Level 3, while SAT has 50% at Level 3. In comparing these percentages, however, it is important to note the great difference in the structure of the frameworks and number of objectives in each framework.

DOK Levels of the Test Items

Panelists assigned each item a DOK rating, independent of any content alignment. Because panelists were not required to reach consensus on the DOK values of items, these ratings were not consensus ratings, and interrater agreement for DOK is addressed in Table 2. The average DOK levels of the NAEP items in the short-form set of 40 items used for the NAEP-to-NAEP study were 2.26 for Panel 1 and 2.20 for Panel 2. The average DOK levels of the NAEP items in

the complete set of 131 items used for the NAEP-to-SAT study were 2.37 for Panel 1 and 2.27 for Panel 2. The average DOK levels for the SAT items in the short-form set of 46 items were 2.50 for Panel 1 and 2.47 for Panel 2. For the SAT items in the complete set of 134 items (67 per form) used in the SAT-to-NAEP study, the average DOK levels were 2.47 for Panel 1 and 2.44 for Panel 2 in Form A, and 2.36 for Panel 1 and 2.34 for Panel 2 in Form F. Thus, the samples appeared representative of the complete pools in terms of DOK. 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 analysis later in this section.

Alignment Results by Sub-Study

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

- Sub-Study 1. NAEP Items (Short Version) to NAEP Framework
- Sub-Study 2. SAT Items (Forms A and F) to NAEP Framework
- Sub-Study 3. SAT Items (Short Version) to SAT Framework
- Sub-Study 4. NAEP Items to SAT Framework

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

In Sub-Study 1, panelists evaluated the alignment between the NAEP items and the NAEP framework. A short-version 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 by each panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 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	44.14	100	43.43	100
Uncodable	0.00	0	0.00	0
Total	44.14		43.43	

For the 40 items, the total mean hits across the two panels were 44.14 and 43.43. 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

Standards	Panel 1			Panel 2		
	Mean Hits	% of Total Hits	% of Hits Adjusted for Uncodable	Mean Hits	% of Total Hits	% of Hits Adjusted for Uncodable
1 - Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension	9.00	20	20	10.14	23	23
2 - Integrate/Interpret: Make complex inferences within and across texts	29.43	67	67	26.86	62	62
3 - Critique/Evaluate: Consider text(s) critically	5.71	13	13	6.43	15	15
Total	44.14	100	100	43.43	100	100

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

All NAEP standards received hits from NAEP items in the short-version subset, with a distribution as indicated in Table 7. Of the three standards, Standard 2, "Integrate/Interpret," received the majority of mean hits in both Panels 1 and 2 (29.43 and 26.86, respectively), making up 67% and 62% of the item set, respectively. Standard 1, "Locate/Recall," had 9.00 mean hits in Panel 1 and 10.14 in Panel 2, for 20% and 23% of total hits, respectively. Standard 3, "Critique/Evaluate," received the fewest mean hits, 5.71 and 6.43 in Panels 1 and 2, or 13% and 15% of the total hits, respectively.

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	3.43	8	6.29	14
	1.2	1.2.a	1	2	0.29	1
		1.2.b	1	2	1	2
		1.2.c	0	0	0	0
		1.2.d	0	0	0	0

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		1.2.e	0	0	0	0
	1.3	1.3.a	1.43	3	1.57	4
		1.3.b	0.14	0	0	0
		1.3.c	2	5	0.71	2
		1.3.d	0	0	0.14	0
		1.3	0	0	0.14	0
2	2.1	2.1.a	0.14	0	0.14	0
		2.1.b	2.43	6	1.57	4
		2.1.c	1	2	1.57	4
		2.1.d	5.14	12	4.43	10
		2.1.e	0	0	0	0
		2.1.f	0.71	2	1.29	3
	2.2	2.2.a	0.86	2	1.14	3
		2.2.b	0.71	2	1	2
		2.2.c	2.57	6	2.86	7
		2.2.d	0.29	1	1.29	3
		2.2.e	0	0	0	0
	2.3	2.3.a	2.43	6	0.43	1
		2.3.b	4.14	9	2.71	6
		2.3.c	0.57	1	0.71	2
		2.3.d	0	0	0	0
		2.3.e	0.14	0	0.14	0
		2.3	0	0	0.14	0
	2.4	2.4.a	8.29	19	7.43	17
3	3.1	3.1.a	3.29	7	2.71	6
		3.1.b	0.57	1	0.86	2
		3.1.c	0	0	0	0
	3.2	3.2.a	0	0	0	0
		3.2.b	0.57	1	1	2
		3.2.c	0	0	0	0
	3.3	3.3.a	0.14	0	0.14	0
		3.3.b	0.86	2	1.14	3
		3.3.c	0.29	1	0.43	1
		3.3.d	0	0	0.14	0

As shown in Table 8, the following three 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"
- 2.4.a, "Determine word meaning as used in context"

Of these three objectives, 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 relatively broad in scope. Objective 2.4.a received the greatest number of hits and the highest percentage (19% and 17%) of total hits to all objectives; this is the only objective addressing vocabulary, and one would expect to see all (or nearly all) vocabulary items mapped to 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.

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"
- 2.1.e, "Describe or analyze how an author uses organizing structures to convey meaning"
- 2.2.e, "Explain how rhythm, rhyme, sound, or form in poetry contribute to meaning"
- 2.3.d, "Distinguish fact from opinion"
- 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.c, "Evaluate a character's conflict, motivations, and decisions"

Two objectives for "Locate/Recall" (Objective 1.2.e, which had no hits, and Objective 1.3.d, which had one hit from one panelist) were objectives about which panelists had raised questions during the pilot study, observing that, in most cases, organizing structures of texts are not explicit but have to be inferred from evidence in the text; panelists had difficulty reconciling the focus of these two objectives 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 an expert reading consultant to the Governing Board, and in this 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.

In addition to the objectives with no hits (listed previously), nine objectives received a mean hit value of less than 1.0. A mean hit value of less than 1.0 indicates that while at least one panelist aligned an item or items to an objective, the objective received fewer than seven total hits across all items and panelists.

- 1.3.b, “Locate or recall the author’s purpose”
- 1.3.d, “Locate or recall organizing structures of texts, such as comparison/contrast, problem/solution, enumeration, etc.”
- 2.1.a, “Describe problem and solution, or cause and effect”
- 2.3.c, “Find evidence in support of an argument”
- 2.3.e, “Determine the importance of information within and across texts”
- 3.1.b, “Analyze, critique, or evaluate the author’s perspective or point of view”
- 3.3.a, “Evaluate the way the author selects language to influence readers”
- 3.3.c, “Determine the quality of counterarguments within and across texts”
- 3.3.d, “Judge the coherence or logic of an argument”

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: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension	9	10.14	100	100	47*	37**	0.76	0.68*
2 - Integrate/Interpret: Make complex inferences within and across texts	29.43	26.86	82	74	58	66	0.67*	0.68*
3 - Critique/Evaluate: Consider text(s) critically	5.71**	6.43	97	100	30**	37**	0.77	0.81

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

Of the 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.” Using the typical WAT threshold value of six mean hits, categorical concurrence was met for Standard 1, “Locate/Recall,” with 9 and 10.14 mean hits to the standard. Categorical concurrence was also met for Standard 2, “Integrate/Interpret,” with 29.43 and 26.86 mean hits to the

standard. Categorical concurrence was met for Standard 3, "Critique/Evaluate," in Panel 2, with 6.43 mean hits to the standard, and nearly met in Panel 1, with 5.71 mean hits to the standard, just below the typical threshold. 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 74%–100% of the items at or above the DOK level of the standard to which they aligned. Still well over the threshold value, the lower DOK consistency for "Integrate/Interpret" (82% and 74%) compared to that of "Locate/Recall" and "Critique/Evaluate" (100%) most likely reflects the greater range of depth of knowledge levels implicit in the language of "Integrate/Interpret." 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) and the summary of main ideas (typically DOK Level 2) to the analysis of theme (DOK Level 3) and 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," corresponds to DOK Level 3 ("Students are encouraged to go beyond the text").

Range of knowledge was met for "Integrate/Interpret," with 58% and 66% of the objectives in this standard covered by NAEP items. Range of knowledge was not met by the 40-item short version for "Locate/Recall" and "Critique/Evaluate," with only 30% to 47% of the objectives in these standards covered across the two panels. For "Locate/Recall," 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 "Critique/Evaluate," 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 "Critique/Evaluate" in both panels and for "Locate/Recall" in Panel 1, and it was weakly met for "Integrate/Interpret." For "Integrate/Interpret," 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 "Locate/Recall" and "Critique/Evaluate," while fewer objectives received hits, the distribution of hits among those objectives was more even. For "Critique/Evaluate," for example, the difference in number of hits per objective averaged approximately 0.5 hits.

Sub-Study 2—SAT Items (Forms A and F) to NAEP Framework

In Sub-Study 2, reviewers evaluated the alignment between the SAT items and the NAEP framework. Two 67-item forms (Forms A and F) of the SAT assessment were aligned with the NAEP framework, for a total of 134 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 coded it to an objective. For an item to be uncodable, all reviewers must have rated 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—SAT Items (Forms A and F) to NAEP Framework
Assessment items = 134 (67 items per form)

	SAT Form A		**SAT Form F**	
	Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
Codable items	67	67	67	67
Uncodable items	0	0	0	0
Total assessment items	67	67	67	67

As shown in Table 10, all SAT 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 by each panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 11. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Seven Reviewers per Panel—SAT Items (Forms A and F) to NAEP Framework
Assessment items = 134 (67 items per form)

	SAT Form A				**SAT Form F**			
	Panel 1		**Panel 2**		**Panel 1**		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**	**Mean Hits**	**Percentage**	**Mean Hits**	**Percentage**
Codable	73.14	100	71.57	100	74.71	100	70.86	100
Uncodable	0.00	0	0.00	0	0.14	0	0.00	0
Total	73.14		71.57		74.86		70.86	

For SAT Form A, the total mean hits were 73.14 for Panel 1 and 71.57 for Panel 2. For SAT Form F, the total mean hits were 74.86 for Panel 1 and 70.86 for Panel 2. These numbers exceed 67 per form because some items were coded to multiple objectives by one or more panelists. One panelist in Panel 1 assigned an uncodable rating to one item.

Table 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—SAT Items (Forms A and F) to NAEP Framework
Assessment items = 134 (67 items per form)

	SAT Form A						SAT Form F					
	Panel 1			Panel 2			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**	**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 . . .	1.00	1	1	2.43	3	3	1.29	2	2	2.57	4	4
2 - Integrate/Interpret: Make complex inferences within and across texts	70.43	96	96	64.00	89	89	71.00	95	95	65.57	93	93
3 - Critique/Evaluate: Consider text(s) critically	1.71	2	2	5.14	7	7	2.43	3	3	2.71	4	4
Total	73.14	100	100	71.57	100	100	74.71	100	100	70.86	100	100

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

All NAEP standards received hits from the SAT items, although the large majority of items were aligned to Standard 2, “Integrate/Interpret.” Of the three standards, “Integrate/Interpret” received the most hits in both panels for both forms (70.43 and 64.00, respectively, for Form A, and 71.00 and 65.57, respectively, for Form F), making up between 89% and 96% of the item set. Standard 1, “Locate/Recall,” received 1.00 and 2.43 mean hits from the two panels for Form A and 1.29 and 2.57 mean hits for Form F, or 1% to 4% of total items. Standard 3, “Critique/Evaluate,” received 1.71 and 5.14 mean hits for Form A and 2.43 and 2.71 mean hits for Form F, or 2% to 7% of the total.

As seen in Table 12, for Form A, Panel 2 assigned 1–3 more mean hits to objectives in “Locate/Recall” and “Critique/Evaluate” than did Panel 1. The discrepancy of greater than five percentage points between the panels for “Critique/Evaluate” in Form A was identified during the study and determined to not be systematic. The discussions indicated a small degree of difference between the panels in the application of “Critique/Evaluate” to SAT “evaluation” items. Although SAT Objective B.1.3 includes “Evaluation,” comments by some panelists in the debrief surveys expressed the judgment that the SAT items rarely required the critical perspective called for in “Critique/Evaluate,” requiring drawing conclusions from a text but not

evaluating the writing. Overall, both panels assigned very few hits to objectives for "Critique/Evaluate."

In comparison with the baseline alignment distribution to the NAEP framework in Sub-Study 1, the SAT items had a range of 20–23 percentage points less emphasis on "Locate/Recall" and 6–12 percentage points less emphasis on "Critique/Evaluate" than did the short-version subset of the NAEP items. The SAT items had a range of 22–34 percentage points greater emphasis on "Integrate/Interpret" than did the short-version subset of the NAEP items.

Table 13. Number and Percentage of Mean Hits to Objectives as Rated by Seven Reviewers per Panel—SAT Items (Forms A and F) to NAEP Framework
Assessment items = 134 (67 items per form)

			SAT Form A				**SAT Form F**			
			Panel 1		**Panel 2**		**Panel 1**		**Panel 2**	
Standards	**Goals**	**Objectives**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
1	1.1	1.1.a	0	0	1.57	2	0.57	1	2.29	3
	1.2	1.2.a	0	0	0	0	0	0	0	0
		1.2.b	0	0	0	0	0	0	0	0
		1.2.c	0	0	0	0	0	0	0	0
		1.2.d	0	0	0	0	0	0	0	0
		1.2.e	0	0	0	0	0	0	0	0
	1.3	1.3.a	0	0	0.14	0	0	0	0.29	0
		1.3.b	0.43	1	0.43	1	0	0	0	0
		1.3.c	0.57	1	0.29	0	0.57	1	0	0
		1.3.d	0	0	0	0	0.14	0	0	0
2	2.1	2.1.a	0.86	1	1.29	2	2	3	3.29	5
		2.1.b	11.57	16	5.43	8	10.43	14	6.57	9
		2.1.c	2	3	2	3	1.86	2	2.29	3
		2.1.d	11.14	15	9.29	13	10.71	14	6.71	9
		2.1.e	1.14	2	1	1	0.57	1	2.71	4
		2.1.f	3.86	5	3.43	5	3	4	3.57	5
		2.1	0	0	0.71	1	0	0	0.14	0
	2.2	2.2.a	2.29	3	3	4	2	3	1.86	3
		2.2.b	0.57	1	0.29	0	0	0	0.29	0
		2.2.c	3	4	4.43	6	1.57	2	3.86	5
		2.2.d	0	0	0.86	1	0	0	1.14	2
		2.2.e	0	0	0	0	0	0	0	0
		2.2	0	0	0.14	0	2.29	3	1.29	2
	2.3	2.3.a	4.57	6	1.29	2	11.71	16	6.14	9
		2.3.b	5.29	7	4.86	7	0.43	1	0.43	1

Standards	Goals	Objectives	SAT Form A				SAT Form F			
			Panel 1		Panel 2		Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits	Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		2.3.c	0.14	0	0.86	1	0	0	0	0
		2.3.d	0	0	0.14	0	0.14	0	1.57	2
		2.3.e	0	0	0.14	0	0.14	0	0	0
	2.4	2.4.a	24	33	24.86	35	24.14	32	23.71	33
3	3.1	3.1.a	0.29	0	0.29	0	0.57	1	0	0
		3.1.b	1	1	2.43	3	0.29	0	0.86	1
		3.1.c	0.14	0	0	0	0	0	0.14	0
	3.2	3.2.a	0	0	0.57	1	0	0	0	0
		3.2.b	0	0	0.14	0	0	0	0	0
		3.2.c	0	0	0	0	0.14	0	0	0
	3.3	3.3.a	0.29	0	0.14	0	0	0	0.14	0
		3.3.b	0	0	0.57	1	0.29	0	0.57	1
		3.3.c	0	0	0	0	1.14	2	0.29	0
		3.3.d	0	0	1	1	0	0	0.71	1

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

As shown in Table 13, the following objectives had the greatest number of mean hits (over five in at least one panel for at least one form):

- 2.1.b, “Compare or connect ideas, perspectives, problems, or situations”
- 2.1.d, “Describe or analyze how an author uses literary devices or text features to convey meaning”
- 2.3.a, “Summarize major ideas”
- 2.3.b, “Draw conclusions and provide supporting information”
- 2.4.a, “Determine word meaning as used in context”

Of these five objectives, Objective 2.4.a received the greatest number of mean hits (at least 23.71) across both panels for both forms. SAT items measure two types of vocabulary skills: the understanding of vocabulary within sentences and understanding of vocabulary in the context of a reading passage. Consistent with a decision rule stating that both types of vocabulary items require the use of context clues, panelists aligned both passage-based and sentence-based vocabulary items to NAEP Objective 2.4.a (see decision rule 1 for the SAT specifications). NAEP Objective 2.1.b has clear parallels to content in SAT Objective B.1.5 (“Application and Analogy”) and to parts of SAT Objective B.1.3 (“Implication and Evaluation”). NAEP Objective 2.1.d also overlaps with SAT Objective B.1.2 (“Rhetorical Strategies”). NAEP Objective 2.3.a, closely related to SAT Objective B.1.1 (“Primary Purpose”), received more mean hits on SAT Form F (11.71 and 6.14) than on Form A (4.57 and 1.29). (Parallel content in the two frameworks is discussed in detail in the Interim Report, included as Appendix B to this report.)

Objectives for Standard 2, "Integrate/Interpret," received the majority of hits from both panels, but the distribution to objectives varied somewhat across panels. Panel 1, for example, assigned more hits to Objective 2.1.b, while Panel 2 assigned more hits to Objective 2.2.c. 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. For example, on Form F, Panel 2 coded some items to Objective 2.2.c ("Interpret mood, tone, or voice") that Panel 1 coded to Objective 2.2.a ("Interpret a character's conflict, motivations, and decisions"). SAT Objective B.1.4 ("Tone and Attitude") includes questions about "the tone *or attitude* of a character in fiction"; questions about a character's "attitude" could involve both tone (NAEP Objective 2.2.c) and motivation (NAEP Objective 2.2.a). In another example, Panel 1 coded some items to Objective 2.1.b ("Compare or connect ideas") that Panel 2 coded to Objective 2.3.b ("Draw conclusions and provide supporting information"). Some items could reasonably be aligned to either or both of these broader NAEP objectives (for example, an item might ask for a conclusion about the connection between two ideas).

Both panels also assigned a small number of hits to two goals for "Integrate/Interpret." Goal 2.1 ("Make complex inferences within and across both literary and informational texts") received fewer than one mean hit on both forms, while Goal 2.2 ("Make complex inferences across literary texts") received 2.29 and 1.29 mean hits on Form F. Panelists assign items to goals when they do not find a match between an item and any objective but believe the item aligns to the broader skill described at the goal level.

As noted previously, objectives for Standard 3, "Critique/Evaluate," received very few hits; only Objective 3.1.b received more than a percentage of one mean hit (2.43) from one panel, on Form A only. The low number of hits to objectives in "Critique/Evaluate" is consistent with differences in the frameworks for the two assessments (discussed in detail in the NAEP–SAT Interim Report, included as Appendix B to this report). SAT Objective B.1.3 includes "Evaluation" as in "evaluate ideas or assumptions in a passage" or "evaluate the relationship between a pair of passages." However, the language of the SAT objective does not suggest the kind of critical perspective characteristic of the third cognitive target described by NAEP, "Critique/Evaluate." According to the NAEP specifications, this type of thinking requires students to "stand back from what they read" and "consider the text critically." Items may "ask students to evaluate the quality of the text as a whole, to determine what is most significant in a passage, or to judge the effectiveness of specific textual features" (NAEP, 2009d, p. 48). Specific NAEP objectives within this dimension include "Judge the author's craft and technique" and "Evaluate the strength and quality of evidence used by the author to support his or her position." None of the SAT objectives specifically call for test takers to take a critical stance, or to judge, critique, or evaluate the *quality* of an author's reasoning or writing.

Objectives for Standard 1, "Locate/Recall," also received very few hits. Only Objective 1.1.a received more than a percentage of one mean hit, from one panel. The SAT framework includes "Literal Comprehension" (Goal B.2), but, unlike NAEP "Locate/Recall," the SAT goal does not specifically refer to the ability to "locate or recall" explicit content from a passage. Rather, "Literal Comprehension questions focus on a small but significant portion of a reading passage and ask what is being said in those lines." Although the broad term "comprehension" could encompass the ability to "locate or recall," it can also include the interpretation of what is stated,

and this is the judgment apparent in the panelists' ratings. It is worth noting that the consensus DOK for SAT Goal B.2 was Level 2, while all the objectives for NAEP "Locate/Recall" were assigned DOK Level 1.

The following NAEP objectives received no hits in the SAT 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"
- 2.2.e, "Explain how rhythm, rhyme, sound, or form in poetry contribute to poetry"

In regard to Objective 2.2.e, it is worth noting that the SAT framework does not include poetry in its reading passages; fiction is the only literary genre represented.

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 14a. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Seven Reviewers per Panel—SAT Items (Form A) to NAEP Framework
Assessment items = 67

	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 . . .	1**	2.43**	100	100	9**	14**	0.69*	0.95
2 - Integrate/Interpret: Make complex inferences within and across texts	70.43	64	86	89	65	72	0.6*	0.58**
3 - Critique/Evaluate: Consider text(s) critically	1.71**	5.14**	94	75	13**	21**	0.83	0.89

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.

Table 14b. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Seven Reviewers per Panel—SAT Items (Form F) to NAEP Framework
Assessment items = 67

	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 . . .	1.29**	2.57**	100	100	10**	11**	0.55**	0.97
2 - Integrate/Interpret: Make complex inferences within and across texts	71	65.57	76	78	61	77	0.57**	0.61*
3 - Critique/Evaluate: Consider text(s) critically	2.43**	2.71**	96	74	17**	17**	0.98	0.91

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 67 SAT assessment items analyzed on each form, all (67 per form, or 134 total) were found to match objectives, or have "hits." The majority of items were coded to Standard 2, "Integrate/Interpret." Using the typical WAT threshold value of six mean hits, categorical concurrence was met for "Integrate/Interpret" on both forms, with 70.43 and 64 mean hits on Form A and 71 and 65.57 mean hits on Form F for Panels 1 and 2, respectively. Categorical concurrence was not met for Standard 1, "Locate/Recall," on either form, with fewer than three mean hits to the standard in both panels. Categorical concurrence was also not met for Standard 3, "Critique/Evaluate," with fewer than six mean hits to the standard on either form in both panels.

In both panels, depth-of-knowledge consistency was met for all three standards. For "Locate/Recall," 100% of the aligned items on both forms were rated at or above the DOK level of the standard. For "Integrate/Interpret," 86% and 89% of aligned items on Form A and 76% and 78% on Form F were rated at or above the DOK level of the standard. For "Critique/Evaluate," 94% and 75% of aligned items on Form A and 96% and 74% on Form F were rated at or above the DOK level of the standard. For DOK consistency in "Critique/Evaluate," a discrepancy between panels of greater than five percentage points was identified. However, cross-panel adjudication revealed this discrepancy to be attributable to the very small number of hits in both panels, and not a systematic difference.

On both forms, range of knowledge was met for "Integrate/Interpret," with 65% and 72% of its objectives hit on Form A and 61% and 77% on Form F. "Locate/Recall" had a range of knowledge of 9% and 14% on Form A and 10% and 11% on Form F. "Critique/Evaluate" had a

range of knowledge of 13% and 21% on Form A and 17% in both panels on Form F. The limited range of knowledge for "Locate/Recall" and "Critique/Evaluate" is to be expected given the low number of mean hits, and is consistent with differences in the frameworks for the two tests. As previously discussed, the SAT framework does not include the skill of critiquing the quality or effectiveness of reading passages, which is a skill included in NAEP "Critique/Evaluate." In addition, all SAT items are multiple choice; the language of many of the objectives for NAEP "Critique/Evaluate" suggests that they would be most effectively measured through constructed-response items (for example, Objectives 3.1.a, "Judge the author's craft and technique," or 3.3.b, "Evaluate the strength and quality of evidence used by the author to support his or her position"). SAT Goal B.2, "Literal Comprehension," is stated more broadly than NAEP "Locate/Recall," referring to comprehension or understanding of explicit content but not to the ability to locate or recall.

The findings regarding balance of representation varied somewhat across panels. Both panels found balance of representation was met for "Critique/Evaluate" on both forms. In other words, although both panels assigned relatively few hits to "Critique/Evaluate," the hits assigned were fairly evenly distributed to those objectives receiving hits (for example, eight of the ten objectives receiving hits received less than one mean hit). For "Locate/Recall," Panel 2 found balance of representation was met on both forms, while Panel 1 found that balance of representation was only weakly met on Form A and not met on Form F. For "Integrate/Interpret," Panel 1 found balance was weakly met on Form A while Panel 2 found that balance was not met on that form. On Form F, Panel 1 found that balance was not met for "Integrate/Interpret," while Panel 2 that found balance was weakly met.

Given that relatively few hits were assigned to objectives for "Locate/Recall" and "Critique/Evaluate" the findings for balance of representation for those standards are probably less meaningful than those for "Integrate/Interpret," which received the great majority (89% or greater for each panel and form) of the total hits. For "Integrate/Interpret," the weak results for balance of representation reflect the fact that the majority of hits to the standard were to four of the 14 objectives receiving hits. These included Objective 2.4.a, the vocabulary objective, which received 24 mean hits, approximately one-third of all mean hits across all standards and objectives and twice as many as any other "Integrate/Interpret" objective. As discussed earlier in relation to Table 13, all four of the objectives receiving the most hits describe skills that are directly paralleled in the SAT framework. They also tend to be among the more broad "Integrate/Interpret" objectives ("Compare or connect ideas," for example). Objectives that received relatively few hits (under three mean hits) were typically more specific than those in the SAT framework and/or describe skills not explicitly addressed in the SAT framework. For example, Objective 2.2.b ("Integrate ideas to determine theme") received less than one mean hit on both forms. Although SAT includes fiction in its reading passages, the framework does not refer to theme or other specifically literary elements of texts. Objective 2.1.c ("Determine unstated assumptions in an argument") received 2 mean hits on Form A and 1.86 and 2.29 mean hits on Form F. This objective describes a more specific application of a broader skill described in the SAT framework ("Implication and Evaluation," including inferring "an author's views").

Sub-Study 3—SAT Items (Short Version) to SAT Framework

In Sub-Study 3, reviewers evaluated the alignment between the SAT Critical Reading items and the SAT reading framework. One 46-item short-version form sampled from the SAT Critical Reading test (Form A) was analyzed. The results of Sub-Study 3 are presented in Tables 15–19.

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

Table 15. Codability of Items as Determined by Items Rated Uncodable by Seven Reviewers per Panel—SAT Items (Short Version) to SAT Framework
Assessment items = 46

	Panel 1	**Panel 2**
Codable items	46	46
Uncodable items	0	0
Total assessment items	46	46

As shown in Table 15, all SAT 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 by each panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 16. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Seven Reviewers per Panel—SAT Items (Short Version) to SAT Framework
Assessment items = 46

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	47.29	100	45.71	99
Uncodable	0.00	0	0.29	1
Total	47.29		46.00	

For the 46 SAT items, the two panels had 47.29 and 46.00 total mean hits. This number exceeds 46 in Panel 1 because some items were coded to multiple objectives by one or more panelists. Two panelists in Panel 2 assigned an uncodable rating to two item(s).

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 Panel 1 in 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—SAT Items (Short Version) to SAT Framework
Assessment items = 46

Standards	Panel 1			Panel 2		
	Mean Hits	% of Total Hits	% of Hits Adjusted for Uncodable	Mean Hits	% of Total Hits	% of Hits Adjusted for Uncodable
A. Sentence Completion	12.71	27	27	10.57	23	23
B. Passage-Based Reading	34.57	73	73	35.14	77	76
Total	47.29	100	100	45.71	100	99

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

All SAT standards received hits from SAT items in the short-version subset. Of the two standards, SAT B ("Passage-Based Reading") received the most mean hits in both panels (34.57 and 35.14), accounting for 73% and 77% of the total hits. SAT A ("Sentence Completion") received 12.71 and 10.57 mean hits (27% and 23% of the total).

Reporting categorical concurrence in terms of mean hits and percentage of hits at a finer grain size, Table 18 displays the numbers and percentages of mean hits to objectives. Percentages for this table are reported as the percentage of total hits.

Table 18. Number and Percentage of Mean Hits to Objectives as Rated by Seven Reviewers per Panel—SAT Items (Short Version) to SAT Framework
Assessment items = 46

Standards	Goals	Objectives	Panel 1		Panel 2	
			Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
A. Sentence Completion		A	12.71	27	10.57	23
B. Passage-Based Reading	B.1	B.1.1	3.71	8	3.57	8
		B.1.2	7.43	16	7.14	16
		B.1.3	8.86	19	9	20
		B.1.4	1.29	3	1.14	2
		B.1.5	5.14	11	4.71	10
	B.2	B.2	4.43	9	3.86	8
	B.3	B.3	3.71	8	5.71	12

As shown in the table, the following objectives received the greatest number of hits (over five mean hits in either or both panels):

- A, "Sentence Completion"
- B.1.2, "Rhetorical Strategies"
- B.1.3, "Implication and Evaluation"

- B.1.5, “Application and Analogy”
- B.3, “Vocabulary in Context”

All of the remaining objectives received hits, with 4.43 and 3.86 mean hits for Goal B.2 (“Literal Comprehension”) and 1.29 and 1.14 mean hits for Objective B.1.4 (“Tone and Attitude”). The percentages of items assigned to each objective were highly consistent across the two panels.

Table 19 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 19. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Seven Reviewers per Panel—SAT Items (Short Version) to SAT Framework
Assessment items = 46[15]

	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**
A. Sentence Completion	15.38	10.57	97	100	100	100	1	1
B. Passage-Based Reading	34.57	35.14	94	91	100	100	0.77	0.77

All alignment criteria were met for both SAT standards. As seen in Table 19, using the typical WAT threshold value of six mean hits, categorical concurrence was met for both standards. In addition, depth-of-knowledge consistency was very high for both SAT A and B.

Both standards met the typical threshold criteria for range of knowledge and balance of representation. Regarding SAT A, however, it is important to remember that, with only one hierarchical level, or one objective in this standard, these criteria can be met with one mean hit. Therefore, range of knowledge and balance of representation are not applicable when referring to SAT A.

[15] The percentages in this table indicate the distribution of total hits. It should be noted that, as shown in Table 16, 1% of the adjusted total hits for SAT items were determined by panelists in Panel 2 to be uncodable to any objective.

Sub-Study 4—NAEP Items to SAT Framework

In Sub-Study 4, panelists evaluated the alignment between the NAEP items and the SAT framework. All 131 NAEP items were analyzed. The results of Sub-Study 4 are presented in Tables 20–24.

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

Table 20. Codability of Items as Determined by Items Rated Uncodable by Seven Reviewers per Panel—NAEP Items to SAT Framework
Assessment items = 131

	Panel 1	**Panel 2**
Codable items	131	131
Uncodable items	0	0
Total assessment items	131	131

As shown in Table 20, all NAEP 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 21 displays the numbers and percentages of mean hits assigned to items by panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 21. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Seven Reviewers per Panel—NAEP Items to SAT Framework
Assessment items = 131

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	144.71	100	132.00	100
Uncodable	0.00	0	0.14	0
Total	144.71		132.14	

For the 131 NAEP items, the two panels had 144.71 and 132.14 total mean hits for Panels 1 and 2, respectively. These numbers exceed 131 because some items were coded to multiple objectives by one or more panelists. In Panel 2, one panelist assigned an uncodable rating to one item.

Table 22 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 22. Categorical Concurrence between Standards and Assessment as Rated by Seven Reviewers per Panel—NAEP Items to SAT Framework
Assessment items = 131

	Panel 1			Panel 2		
Standards	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
A. Sentence Completion	0.43	0	0	0.00	0	0
B. Passage-Based Reading	144.29	100	100	132.00	100	100
Total	144.71	100	100	132.00	100	100

Of the two SAT standards, only SAT B, "Passage-Based Reading," received a non-negligible number of hits from NAEP items. "Passage-Based Reading" received 100% of the total hits (144.29 in Panel 1 and 132.00 in Panel 2). SAT A, "Sentence Completion," received less than one mean hit, or 0% of total hits.

In comparison with the baseline alignment of short-version SAT items to the SAT framework in Sub-Study 3, the NAEP items had virtually no emphasis on "Sentence Completion" (that is, 23–27 percentage points less emphasis than the short-version SAT), and 23–27 percentage points greater emphasis on "Passage-Based Reading" than did the short-version SAT.

Table 23. Number and Percentage of Mean Hits to Objectives as Rated by Seven Reviewers per Panel— NAEP Items to SAT Framework
Assessment items = 131

			Panel 1		Panel 2	
Standards	**Goals**	**Objectives**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
A. Sentence Completion		A	0.43	0	0	0
B. Passage-Based Reading		B	0	0	0.14	0
	B.1	B.1	0.86	1	0.29	0
		B.1.1	10.57	7	6.43	5
		B.1.2	17.14	12	21.29	16
		B.1.3	41.86	29	32.57	25
		B.1.4	1.57	1	2	2
		B.1.5	10.57	7	8.14	6
	B.2	B.2	38.29	26	36.29	27
	B.3	B.3	23.43	16	24.86	19

As shown in Table 23, all objectives for SAT B, "Passage-Based Reading," received one or more hits from each panel. The following objectives received the greatest number of hits (over five mean hits in both panels):

- B.1.1, "Primary Purpose"
- B.1.2, "Rhetorical Strategies"
- B.1.3, "Implication and Evaluation"
- B.1.5, "Application and Analogy"
- B.2, "Literal Comprehension"
- B.3, "Vocabulary in Context"

All of these objectives cover content also covered in the NAEP framework. In addition, because the SAT objectives are broader and more general in scope than the NAEP objectives, most of the SAT objectives overlap with multiple NAEP objectives. SAT Objective B.1.1 ("Primary Purpose"), for example, overlaps with NAEP Objectives 1.3.b ("Locate or recall author's purpose"), 2.1.f ("Describe or analyze author's purpose"), and 2.3.a ("Summarize main ideas"). SAT Objective B.1.3 overlaps at the broad level of making inferences to interpret "implications" with NAEP Standard 2, "Integrate/Interpret," and all of its objectives; because it also includes "evaluation," SAT Objective B.1.3 may also be considered to have some degree of overlap with objectives in NAEP Standard 3, "Critique/Evaluate." The only SAT objective to receive fewer than five hits, Objective B.1.4, is also the most specific; it focuses on understanding the "Tone and Attitude" of an author or character in a passage. Unlike most of the SAT objectives, it has only one direct parallel in the NAEP framework, to NAEP Objective 2.2.a ("Interpret mood, tone, or voice").

SAT A, "Sentence Completion," was the only SAT objective to receive no hits. Although NAEP includes items assessing understanding of vocabulary, it does not address vocabulary in the context of single sentences. All NAEP vocabulary items are passage-based.

Panelists may assign an item to the more general standard or goal level if they do not find a match to an objective. Panel 2 had 0.14 mean hits (0% of total hits) assigned to the standard "Passage-Based Reading," and Panel 1 had 0.86 mean hits (1% of total hits) assigned to SAT Goal B.1 ("Extended Reasoning").

Table 24 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 24. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Seven Reviewers per Panel—NAEP Items to SAT Framework
Assessment items = 131

Objectives	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)	
	Panel 1	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
A. Sentence Completion	0.43**	0**	67	0**	14**	0**	0.14**	0**
B. Passage-Based Reading	144.29	132	87	78	100	100	0.67*	0.68*

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, all were found to match SAT objectives. As previously noted, SAT A, "Sentence Completion," received less than one mean hit; no alignment criteria were met for this standard and objective. As previously noted, the lack of coverage of "Sentence Completion" by NAEP items is the major difference between the distributions of SAT items and NAEP items to the SAT framework. All alignment criteria were met for SAT B, "Passage-Based Reading." Balance of representation was the only criterion to be weakly met; this result reflects the uneven distribution of NAEP items to SAT objectives for "Passage-Based Reading." Of the eight objectives receiving hits, for example, two of the broadest, Objective B.1.3 ("Implication and Evaluation") and Goal B.2 ("Literal Comprehension"), received over 50% of the total hits.

As shown in Table 24, Panel 1 assigned 144.29 mean hits overall while Panel 2 assigned 132. This reflects the fact that Panel 1 assigned more items to multiple objectives than did Panel 2. The great majority of items coded by Panel 1 to two objectives were assigned to both Objective B.1.2 ("Rhetorical Strategies") and Objective B.1.3 ("Implication and Evaluation"). (This was true for the double codes assigned in both panels, although Panel 1 assigned more items to both of these objectives.) SAT Objective B.1.2 items focus on a specific element of a passage and ask "why this particular element is present or what purpose it serves." SAT Objective B.1.3 asks about what an element of a passage "suggests or what can be inferred about the author's views." Panelists found that some NAEP items addressed both of these objectives.

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 25 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 25. 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	3	**4**	0	0	2	**5**
Q6. introduce NAEP/SAT?	0	2	2	**3**	0	2	2	**3**
Q7. prepare you to under-stand DOK levels?	0	0	**5**	2	0	0	**4**	3
Q8. prepare you for the consensus process?	0	1	**3**	**3**	0	0	2	**5**
Q9. prepare you to use the WAT system?	0	1	**4**	2	0	1	**3**	**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	0	3	**4**	0	0	3	**4**
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	**7**	

As shown in Table 25, all panelists across the two panels reported that the introductory session either adequately or very well explained the purposes of the study and prepared them for understanding the DOK levels. All but four (29%) panelists reported that the introductory session introduced the NAEP and SAT assessments adequately or very well; these four panelists, two from Panel 1 and two from Panel 2, responded that the training only somewhat adequately introduced the assessments. In addition, all but one Panel 1 panelist responded that they felt the introductory training prepared them adequately or very well for the discussion process that led to agreement on DOK levels for NAEP objectives across the two panels; the remaining panelist commented that additional, less ambiguous, examples for all the DOK levels would have been helpful in the training. Most of the panelists also felt adequately or very well prepared to use the WAT system, with only two (14%) responding that they felt somewhat prepared. Neither of these two panelists provided further information about their preparation to use the WAT system. 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, all panelists on both panels reported feeling either comfortable or very comfortable. In addition, all the panelists on both panels felt that the group facilitators managed the discussions that led to agreement on DOK levels for NAEP objectives across the two panels very well.

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, one panelist requested information on how the panelists were chosen. Recommendations for improving the alignment training were primarily requests for examples of items for each of the DOK levels and for anchor documents that would guide the coding process. Additionally, panelists reiterated the request for additional examples, and one panelist requested more time for practice of the coding process. When asked what aspects of the day went particularly well, 86% (12) of the 14 panelists mentioned the discussion process; other responses included the examples (7%, or 1) and the practice coding session (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 SAT framework objectives. At the end of the day, 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 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. One panelist in Panel 1 did not complete this questionnaire; therefore, only six responses for Panel 1 are reported.

Table 26. Panelist Responses to Day 2 Training and Process Evaluation Questionnaire

	Panel 1 (n=6)				**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	2	**3**	0	0	**5**	2
Q5. prepare you for the alignment (coding) process?	0	**2**	**2**	**2**	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	1		**5**	0	0		**7**

Overall, the majority of panelists reported feeling adequately or very well prepared to assign DOK levels to items and to code items to objectives. However, one panelist from Panel 1 reported feeling only somewhat prepared to assign DOK levels to items, and three (21%) panelists (two from Panel 1 and one from Panel 2) reported feeling only somewhat prepared to align test items to objectives. Of these three panelists, one panelist from Panel 1 indicated that he/she had initially misunderstood the instructions regarding primary and secondary coding and recommended reviewing the manual instructions at the beginning of the training. This panelist responded feeling comfortable with training and alignment activities on Day 1, however. The other panelist from Panel 1 suggested that having additional completed examples would have been helpful, and the panelist from Panel 2 mentioned that the explanations were not always clear and that sometimes the large group setting was confusing. All three of these panelists were monitored throughout the week to ensure their understanding of the process and their ability to effectively complete all alignment tasks. By the end of Day 3, only one of the panelists on Panel 1 who felt somewhat prepared for both assigning DOK levels to test items and the alignment process on Day 2 still responded that way on the questionnaire.

Panelists generally felt that the day's within-panel discussions regarding alignment codes[16] were facilitated very well, with only one Panel 1 panelist reporting that the facilitator did so moderately well. This was one of the panelists who felt only somewhat comfortable with the alignment process on Day 2; this person recommended providing more clarity in directions and on the relationships among the alignment workshop tasks.

Recommendations for improving the alignment process or requests for more information included providing more examples, more clarity in the directions and explanation, further explanation of the differences between the DOK and objective coding, and more practice time. Ten (77%) of the 13 panelists who responded about activities that went particularly well

[16] 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.

commented on the value of within-panel discussions regarding alignment codes, with additional positive comments pertaining to the value of coding the DOK and alignment to objectives at the same time and the skills of the facilitators. When asked to identify areas in which they felt unprepared, two panelists expressed confusion between the coding of DOK and alignment to the objectives.

Day 3 Process Evaluation

The third day of the study comprised mapping of both SAT and NAEP items to the SAT framework. At the end of the day, panelists were administered a process evaluation questionnaire that solicited feedback on these alignment activities. Table 27 shows results for selected-response questions 4, 5, and 6, by panel. Numbers in bold font represent the highest number of responses for each question, by panel. One panelist in Panel 2 did not complete this questionnaire; therefore, only six responses for Panel 2 are reported.

Table 27. Panelist Responses to Day 3 Process Evaluation Questionnaire

	Panel 1 (n=7)				**Panel 2** (n=6)			
How comfortable do you feel…	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**
Q4. assigning DOK levels to test items?	0	1	**3**	**3**	0	0	2	**4**
Q5. aligning test items to objectives?	0	1	**5**	1	0	1	**4**	1
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		**6**

On Day 3, all but one panelist (in Panel 1) felt comfortable or very comfortable assigning DOK levels to test items, and all but two panelists (one in Panel 1 and one in Panel 2) felt comfortable or very comfortable aligning test items to objectives. These remaining panelists reported feeling somewhat comfortable with these tasks, and commented that more time spent working with examples in the large group would have made them more comfortable with the tasks. As on Day 2, all but one of the panelists felt that the facilitator managed the day's within-panel discussions very well.

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. Two of the nine panelists who responded would have liked to have seen additional examples, and one additional panelist recommended allowing more time for discussion. One panelist also commented that he/she would like to see the final report. When asked what activities went particularly well, panelists expressed appreciation for examples and opportunities to discuss and practice. When asked what could have been done differently to improve the day's activities, the majority (78%, or seven of the nine responses to this question) were positive, indicating that the alignment activities went well. One comment suggested a mechanism for

limiting panelists' discussion about changing assessment items and/or passages, while another panelist recommended splitting alignment coding into two sessions.

Day 4 Process Evaluation

The remaining alignment activities—primarily coding SAT 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 28 shows results for selected-response questions 4, 5, and 6, by panel. Numbers in bold font represent the highest number of responses for each question, by panel. One panelist in each panel did not complete this questionnaire; therefore, only six responses are reported for each panel.

Table 28. Panelist Responses to Day 4 Process Evaluation Questionnaire

	Panel 1 (n=6)				**Panel 2** (n=6)			
How comfortable do you feel…	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**
Q4. assigning DOK levels to test items?	0	1	2	**3**	0	0	**4**	2
Q5. aligning test items to objectives?	0	**2***	1*	**2***	0	1	**5**	0
How well did your facilitator…	**Not Well**	**Moderately Well**	**Very Well**		**Not Well**	**Moderately Well**	**Very Well**	
Q6. facilitate today's consensus process?	0	1	**5**		0	0	**6**	

*n=5 for this question.

The pattern that emerged on both Day 2 and Day 3 of the study—with the majority of the panelists feeling comfortable or very comfortable with the DOK and alignment coding process—continued in Day 4. One panelist in Panel 1 felt somewhat comfortable both assigning DOK levels to test items and aligning test items to objectives; he/she reported feeling that there was some interference between the two processes, which are both "tough mental work." An additional panelist from Panel 1 who felt somewhat comfortable aligning test items to objectives would have liked more preparation for aligning SAT items to the NAEP framework. Finally, one panelist in Panel 2—the same panelist who responded this way to this question on the Day 3 questionnaire—reported feeling only somewhat comfortable aligning test items to objectives, indicating that aligning SAT to NAEP was challenging and repeating the recommendation from Day 3 to provide examples that could be done with the whole group prior to the independent alignment process. As on both Days 2 and 3, all but one of the panelists felt that the facilitator facilitated the day's within-panel discussions very well.

Panelists were asked to provide recommendations for improving the alignment process, to record requests for more information, and to specify activities they felt unprepared for; two panelists suggested providing precoded examples to review as a group prior to the independent coding and within-panel discussions. 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, as well as for the skillful facilitation of the panels.

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 potential threats to the reliability of panelist alignment codes and to identify deficiencies in training or workshop logistics that could be addressed for future alignment studies. Table 29 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 29. Panelist Responses to End-of-Study Evaluation 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	1	**3**	**3**	0	0	3	**4**
Q7. for the consensus process?	0	1	2	**4**	0	0	2	**5**
Q8. for the alignment (coding) process?	0	2	**3**	2	0	1	**3**	**3**
How comfortable did you feel…	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**	**Uncom-fortable**	**Some-what**	**Com-fortable**	**Very Com-fortable**
Q5. assigning DOK levels to objectives?	0	1	2	**4**	0	0	1	**6**
Q6. assigning DOK levels to test items?	0	1	1	**5**	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	**3**	**3**	1	0	**4**	1	2
Q10. on-site training and coding materials?	0	0	2	**5**	0	0	1	**6**
How qualified was your panel…	**Not Quali-fied**	**Some-what Quali-fied**	**Ade-quately Quali-fied**	**Very Quali-fied**	**Not Quali-fied**	**Some-what Quali-fied**	**Ade-quately Quali-fied**	**Very Quali-fied**
Q11. to conduct this type of alignment?	0	0	**4**	3	0	0	**4**	3
How easy was it…	**Not Easy**	**Some-what Easy**	**Ade-quately Easy**	**Very Easy**	**Not Easy**	**Some-what Easy**	**Ade-quately Easy**	**Very Easy**
Q15. to use the WAT for the alignment process?	0	1	1	**5**	0	0	2	**5**

On the Day 1 training and process evaluation questionnaire, all 14 panelists reported that the training prepared them either adequately or very well to understand DOK levels, although, once they actually started assigning DOK levels to items, one panelist reported feeling only somewhat comfortable with this process. The panelists' level of comfort with the within-panel discussion and alignment process remained fairly constant over the course of the week, as evidenced by the responses. By the end of the week, all but one panelist felt adequately or very well prepared for understanding the DOK levels, while all but three (21%) panelists felt adequately or very well prepared for the alignment process. As noted previously, one of the panelists in Panel 1 noted that more training with examples during the earlier part of the week might have been helpful, although this panelist also expressed feeling more comfortable with the entire process overall on Day 5 than on prior days.

All 14 of the panelists responded that their panels were either adequately qualified or very qualified to conduct the alignment activities. However, when asked what could have been done to improve the composition of the panel, one panelist commented that some of the other panelists had difficulty following directions and staying on task, while another commented that those with higher education backgrounds in English may have had more difficulty during the process than those with some background in education and/or reading. The remaining comments were positive, with four panelists specifically reporting that the panel was effective and qualified and that it functioned very well.

All panelists reported that the training and coding materials provided during the alignment workshop were useful, while only half of the panelists (four in Panel 1 and three in Panel 2) felt that way about information provided in advance of the study. Thirteen of the fourteen panelists responded positively to the WAT, although some commented about minor technical problems (e.g., timing out) or recommended adding additional functionality, such as using the number key for data entry or removing the "go to item" button on the drop-down menu bar.

Overall, panelists found the facilitators to be effective and reported that facilitators provided strong facilitation for the groups. All 14 panelists reported that the facilitators were effective, fair, and pleasant, providing comprehensive information and demonstrating experience. However, one panelist mentioned that adjudication discussions were at times too fast and that there was not enough time to process all the information. Another panelist, who had reported feeling comfortable with the alignment process early in the week but felt less confident with cross-framework alignment activities, reported that sometimes the rules for adjudication were unclear.

More substantive issues were addressed via open-ended questions. Regarding the utility of the alignment criteria in capturing aspects of each assessment (Question 14), panelist responses focused on either the alignment criteria or the frameworks used in the study. Those who discussed the criteria reported them to be useful, particularly for within-assessment alignment. One panelist noted that, coupled with the decision rules, the criteria were very helpful, with another panelist commenting that examples were useful.

All of the panelists reported that the alignment process did an adequate job of capturing the content *similarities* between the assessments, although one panelist suggested that the process of alignment to frameworks might not, in and of itself, reveal more qualitative differences between

the assessments. All panelists reported that the alignment process effectively captured content *differences* between the assessments. One panelist suggested that the use of notes enhanced the process, while another panelist reported that the SAT content was more challenging to code.

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. Three panelists noted that the SAT framework appears to have a more narrow focus, while the NAEP framework appears to be written for a wider range of skills. Other panelists reported differences in item type (multiple choice versus open or constructed response), as well as NAEP's focus on life/workplace skills compared to the focus on higher education requirements assessed by SAT. Two panelists noted that NAEP included more items at the DOK 1 and DOK 2 levels, while SAT included more items with higher DOK ratings.

When asked about the facilities for the alignment workshop, panelists felt that they were suitable. Table 30 shows panelist responses to this question, by panel. Numbers in bold font represent the highest number of responses for each question, by panel. Only six panelists from Panel 2 completed this portion of the questionnaire.

Table 30. Panelist Responses Regarding Adequacy of Facilities

	Panel 1 (n=7)				**Panel 2** (n=6)			
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	**5**	0	0	2	**4**
Computers and equipment	0	0	0	**7**	0	0	2	**4**
Meals and breaks	0	0	2	**5**	0	0	1	**5**
Sleeping rooms	0	0	0	**7**	0	0	0	**6**

Across both panels, all the panelists who responded reported that the meeting rooms, computers and equipment, meals and breaks, and sleeping rooms were either adequately or very suitable for this type of alignment workshop, although two panelists recommended providing eating rooms separate from meeting rooms and one panelist expressed disappointment that free Internet access was not provided in the sleeping rooms.

Overall, panelists enjoyed the opportunity to participate in this study and felt that the meetings were well planned and well organized.

V. Summary and Conclusions

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

Summary of Overlap of Content Alignment

Table 31 shows the overlap of content alignment of each assessment to its own and the other assessment's framework in terms of the percentages of total hits.

Table 31. Summary of the Overlap of Content Alignment between NAEP and SAT Items and the NAEP Framework and SAT Framework at the Standard Level

	NAEP Items (40 items)		**SAT (A) Items** (67 Items)		**SAT (F) Items** (67 Items)	
	Panel 1	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
NAEP Framework	**% of Total Hits**		**% of Total Hits**		**% of Total Hits**	
1 - Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension	20	23	1	3	2	4
2 - Integrate/Interpret: Make complex inferences within and across texts	67	62	96	89	95	93
3 - Critique/Evaluate: Consider text(s) critically	13	15	2	7	3	4
	NAEP Items (131 items)		**SAT Items** (46 items)			
	Panel 1	**Panel 2**	**Panel 1**		**Panel 2**	
SAT Specifications	**% of Total Hits**		**% of Total Hits**		**% of Total Hits**	
A. Sentence Completion	0	0	27		23	
B. Passage-Based Reading	100	100	73		77	

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

NAEP items on the short-version subset of 40 items were found to assess all NAEP standards (1–3). NAEP items in the full pool of 131 items were found to assess one of the two SAT standards (SAT B, "Passage-Based Reading"). SAT A, "Sentence Completion," was not assessed on NAEP; this skill is also not addressed in the NAEP framework.

SAT items from the short-version subset of 46 items were found to assess both of the SAT standards ("Sentence Completion" and "Passage-Based Reading"). SAT items from the two forms of 67 items each (134 total) were found to assess all three of the NAEP standards (Standard 1, "Locate/Recall"; Standard 2, "Integrate/Interpret"; and Standard 3,

"Critique/Evaluate"). However, the majority of SAT items were found to assess "Integrate/Interpret," while "Locate/Recall" and "Critique/Evaluate" each received minimal coverage.

With regard to alignment to the NAEP framework, both SAT items and NAEP items (short version) had a majority of hits to "Integrate/Interpret" (67% and 62% for NAEP; 96% and 89% for SAT Form A; 95% and 93% for SAT Form F). The remaining NAEP item alignments were divided between "Locate/Recall" (20% and 23%) and "Critique/Evaluate" (13% and 15%). SAT had very limited coverage of "Locate/Recall" (less than 5% of total hits in either form) and "Critique/Evaluate" (2% and 7% of total hits for Form A and 3% and 4% of total hits for Form F).

With regard to the SAT framework, all (100%) of the NAEP hits and the majority (73% and 77%) of the SAT (short version) hits were to "Passage-Based Reading." The remaining approximately one-quarter of the SAT total hits were to "Sentence Completion." Although the NAEP items were not found to address "Sentence Completion," it is important to note that NAEP does assess vocabulary knowledge, a primary focus of "Sentence Completion." In fact, as seen in the following table, 19% and 17% of the total hits from NAEP items were assigned to the NAEP vocabulary objective, 2.4.a. The difference, as noted elsewhere, is a matter both of item format (sentence completion versus passage-based questions) and of the type of application of vocabulary knowledge. Although the SAT sentence-completion items do provide some content clues, the role of context is limited by the single-sentence format; the NAEP passage-based vocabulary items are intended to measure "meaning vocabulary," that is, "the application of one's understanding of word meanings to passage comprehension" (National Assessment Governing Board, 2008, p. 32).

Overall, the NAEP items covered all the NAEP standards and the SAT "Passage-Based Reading" standard; the SAT items covered the SAT standards and primarily covered NAEP "Integrate/Interpret."

Overlap in content alignment to the NAEP framework can also be examined at the more finely grained objective level. Table 32 shows the overlap of alignment of each assessment to the NAEP framework in terms of the percentages of total hits.

Table 32. Summary of the Overlap of Content Alignment between NAEP and SAT Items and the NAEP Framework at the Objective Level

NAEP Framework			**NAEP Items** (40 items)		**SAT Items (Form A)** (67 items)		**SAT Items (Form F)** (67 items)	
			Panel 1	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
Standard	**Goal**	**Objective**	**% of Total Hits**		**% of Total Hits**		**% of Total Hits**	
1	1.1	1.1.a	8	14	0	2	1	3
	1.2	1.2.a	2	1	0	0	0	0
		1.2.b	2	2	0	0	0	0
		1.2.c	0	0	0	0	0	0

NAEP Framework			NAEP Items (40 items)		SAT Items (Form A) (67 items)		SAT Items (Form F) (67 items)	
			Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
Standard	Goal	Objective	% of Total Hits		% of Total Hits		% of Total Hits	
		1.2.d	0	0	0	0	0	0
		1.2.e	0	0	0	0	0	0
	1.3	1.3	0	0	0	0	0	0
		1.3.a	3	4	0	0	0	0
		1.3.b	0	0	1	1	0	0
		1.3.c	5	2	1	0	1	0
		1.3.d	0	0	0	0	0	0
2	2.1	2.1	0	0	0	1	0	0
		2.1.a	0	0	1	2	3	5
		2.1.b	6	4	16	8	14	9
		2.1.c	2	4	3	3	2	3
		2.1.d	12	10	15	13	14	9
		2.1.e	0	0	2	1	1	4
		2.1.f	2	3	5	5	4	5
	2.2	2.2	0	0	0	0	0	0
		2.2.a	2	3	3	4	3	3
		2.2.b	2	2	1	0	0	0
		2.2.c	6	7	4	6	2	5
		2.2.d	1	3	0	1	0	2
		2.2.e	0	0	0	0	0	0
	2.3	2.3	0	0	0	0	0	0
		2.3.a	6	1	6	2	3	2
		2.3.b	9	6	7	7	16	9
		2.3.c	1	2	0	1	1	1
		2.3.d	0	0	0	0	0	0
		2.3.e	0	0	0	0	0	2
	2.4	2.4.a	19	17	33	35	32	33
3	3.1	3.1.a	7	6	0	0	1	0
		3.1.b	1	2	1	3	0	1
		3.1.c	0	0	0	0	0	0

NAEP Framework			NAEP Items (40 items)		SAT Items (Form A) (67 items)		SAT Items (Form F) (67 items)	
			Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
Standard	**Goal**	**Objective**	**% of Total Hits**		**% of Total Hits**		**% of Total Hits**	
	3.2	3.2.a	0	0	0	1	0	0
		3.2.b	1	2	0	0	0	0
		3.2.c	0	0	0	0	0	0
	3.3	3.3.a	0	0	0	0	0	0
		3.3.b	2	3	0	1	0	1
		3.3.c	1	1	0	0	2	0
		3.3.d	0	0	0	1	0	1

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

As shown in Table 32, within the strong standard-level overlap at NAEP Standard 2, "Integrate/Interpret," 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.

In both tests, most of the items assigned to "Integrate/Interpret" tended to cluster on a small number of objectives. For example, on both tests, the following objectives received a higher percentage of hits (over five mean hits, across all forms) than did other Standard 2 objectives:

- 2.1.b, "Compare or connect ideas, perspectives, problems, or situations"
- 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"

All four of the NAEP objectives listed above have close parallels to objectives in the SAT framework. Two of the four, NAEP 2.1.d and NAEP 2.3.b, received similar percentages of hits on both tests, suggesting a similar emphasis on these broad skills in both assessments. Both tests also had relatively high percentages of their overall hits to Objective 2.4.a; however, the SAT had approximately twice as many hits to 2.4.a, consistent with the higher overall percentage of vocabulary items (approximately 35% total, 28% Sentence Completion) called for in the SAT specifications. The percentages of NAEP hits to 2.4.a are also consistent with the target for vocabulary in the NAEP Framework, which calls for (approximately) two of ten items per passage, or 20%, to address vocabulary in context. Overall, compared to the NAEP, the SAT has a higher percentage of items addressing vocabulary but a much lower percentage of items addressing vocabulary in the context of a reading passage.

The SAT items also had roughly double the percentage of hits to Objective 2.1.b. This is the only "Integrate/Interpret" objective whose text explicitly addresses comparison or connection within or across texts. However, as stated in the standard-level descriptions, all of the NAEP standards,

goals, and objectives apply both within and *across* texts. For example, NAEP items that primarily assess Objective 2.3.a ("Summarize major ideas") or Objective 2.2.c ("Interpret a character's conflicts, motivations and decisions") could require applying those skills to multiple passages. For that reason, the higher percentage of SAT items assigned to Objective 2.1.b does not necessarily imply a greater emphasis on comparison within and across texts in SAT. It more likely reflects differences in the level of specificity of the two frameworks; two of the seven SAT "Passage-Based Reading" objectives, B.1.3 and B.1.5, call for students to make comparisons or connections between ideas within and across passages. NAEP includes an objective, 2.1.b, focused specifically on comparisons and connections ("Compare or connect ideas, perspectives, problems, or situations") but also many other more specific objectives that may require comparison or connection when applied across two texts.

Several other "Integrate/Interpret" objectives received similar percentages of hits on both tests:

- 2.1.c, "Determine unstated assumptions in an argument"
- 2.2.a, "Interpret mood, tone, or voice"
- 2.2.c, "Interpret a character's conflicts, motivations, and decisions"
- 2.3.a, "Summarize major ideas"

The results suggest that the skills described in these objectives are represented in roughly similar proportions on both tests, keeping in mind the variations in percentages across SAT forms.

Perhaps the greatest variation between the two tests is found in the percentages of hits to objectives for "Locate/Recall" and "Critique/Evaluate." Very few SAT items were coded to any objective in "Locate/Recall"; only Objective 1.1.a ("Locate or recall specific information such as definitions, facts, and supporting details in texts or graphics") received more than 1% of total hits. In contrast, 8% and 14% of the total NAEP hits were to Objective 1.1.a, and four other "Locate/Recall" objectives received hits in percentages ranging from 2–5% of NAEP items. Although the SAT framework includes an objective for "Literal Comprehension," the findings suggest that the SAT items generally go beyond the skills of locating or recalling information. It may be worth noting, however, that with the exception of Objective 1.1.a, "Locate/Recall" objectives received relatively low percentages of hits from the NAEP (short form) items as well.

Compared to the NAEP items, the SAT items also had smaller percentages of hits to objectives for "Critique/Evaluate." Objective 3.1.a ("Judge the author's craft and technique"), for example, received about 7% of the total NAEP short-version hits but only 1% of the total SAT hits for one panel in Form F only. Objective 3.3.b ("Evaluate the strength and quality of evidence used by the author to support his position") received 2% and 3% of the total NAEP hits but only 1% of SAT hits from one panel only for each form. The percentages of hits to the remaining "Critique/Evaluate" objectives were relatively low (0–3%) for both tests. No hits were assigned to Objectives 3.1.c, 3.2.c, or 3.3.a on either test.

Overall, the results for "Critique/Evaluate" objectives suggest that NAEP differs from SAT in its emphasis on critical evaluation, particularly on the evaluation of the quality or effectiveness of an author's writing.

The above findings appear consistent with the language used in the SAT specifications. The word "evaluate" appears just once in a single sentence in SAT B.1.3, in reference to items that might "ask the test taker to evaluate ideas or assumptions in a passage, or to evaluate the relationship between a pair of passages." Beyond the use of the word "evaluate" (applied to ideas, not to author's craft), the language of SAT B.1.3 does not specify the kind of critical judgment in relation to a text called for in NAEP "Critique/Evaluate" objectives (the "critique" of an author's logic, use of evidence, or rhetorical devices). SAT B.1.2, "Rhetorical Strategies," asks students to identify the purpose of an author's rhetorical strategies, but not to evaluate their quality or effectiveness.

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

SAT Framework			NAEP Items (131 items)		SAT Items (46 items)	
			Panel 1	Panel 2	Panel 1	Panel 2
Standard	Goals	Objectives	% of Total Hits	% of Total Hits	% of Total Hits	% of Total Hits
A		A	0	0	27	23
B	B.1	B	0	0	0	0
		B.1	1	0	0	0
		B.1.1	7	5	8	8
		B.1.2	12	16	16	16
		B.1.3	29	25	19	20
		B.1.4	1	2	3	2
		B.1.5	7	6	11	10
	B.2	B.2	26	27	9	8
	B.3	B.3	16	19	8	12

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

In relation to the SAT framework, both tests had a majority of hits to objectives for SAT B, "Passage-Based Reading"; for the NAEP items, 100% of hits were to "Passage-Based Reading" (NAEP does not address the content in SAT A, "Sentence Completion"). As seen in Table 33, both tests had similar percentages of hits to many of the "Passage-Based Reading" objectives:

- B.1.1, "Primary Purpose"
- B.1.2, "Rhetorical Strategies"
- B.1.4, "Tone and Attitude"

Panelists found slightly more emphasis on Objective B.1.5 ("Application and Analogy") in the SAT (short version) items and slightly more emphasis on Objective B.1.3 ("Implication and Evaluation") in the NAEP items. This difference is not surprising; as previously discussed, two of the seven SAT objectives include making comparisons between ideas or passages, while *all* NAEP Standard 2, "Integrate/Interpret," objectives require making inferences and interpreting implications. The percentage of hits to Goal B.3 ("Vocabulary in Context") was also higher for

NAEP; all NAEP vocabulary items are passage-based, while the SAT includes far fewer passage-based items, primarily addressing vocabulary through its "Sentence Completion" items. All of the SAT "Passage-Based Reading" objectives above include skills and knowledge also covered in the NAEP framework; although there are some variations, the overall results indicate that the two tests give comparable emphasis to the skills described in SAT B.1.1, B.1.2, and B.1.4.

One notable difference between the two tests is the much higher percentage of total hits for SAT Goal B.2 ("Literal Comprehension") from NAEP items compared to the SAT short-version items. This is similar to the difference seen in the percentages of hits in each test to NAEP Standard 1. Although the language of "Locate/Recall" and that of SAT Goal B.2 differ somewhat, both encompass the ability to "locate or recall" explicitly stated content, using simple inferences as needed; the more general SAT objective would appear to include basic interpretation of literal content. The different results in relation to both frameworks suggests a greater emphasis on literal comprehension in NAEP.

Categorical Concurrence

For alignment to the NAEP framework, the NAEP short-version items were found to meet categorical concurrence for all three standards ("Locate/Recall," "Integrate/Interpret," and "Critique/Evaluate"). The SAT items met categorical concurrence for "Integrate/Interpret," but not for "Locate/Recall" or "Critique/Evaluate."

For alignment to the SAT framework, the SAT short-version items were found to meet categorical concurrence for both standards ("Sentence Completion" and "Passage-Based Reading"). The NAEP items met categorical concurrence for "Passage-Based Reading," but not for "Sentence Completion."

Depth-of-Knowledge Consistency

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

For alignment to the SAT framework, the SAT short-version items were found to meet depth-of-knowledge consistency in both standards. The NAEP items met depth-of-knowledge consistency for "Passage-Based Reading," the only SAT standard to which they were aligned.

Range-of-Knowledge Correspondence

For alignment to the NAEP framework, for both NAEP (short version) and SAT items, only "Integrate/Interpret" met the typical WAT range of knowledge threshold criterion. That is, only "Integrate/Interpret" had at least 50% of its objectives receive hits from items in both tests. For the NAEP items, the two panels differed on "Locate/Recall": Panel 1 found that "Locate/Recall" had a weak range of knowledge, with 47% of its objectives receiving hits from NAEP items, while Panel 2 found that it did not show range, with only 37% of its objectives hit. For the SAT

items, both panels found that "Locate/Recall" did not have a range of knowledge, with less than 15% of its objectives receiving hits on both forms. Both panels also found that "Critique/Evaluate" lacked range of knowledge on both tests: 30% and 37% of the objectives in this standard received hits from NAEP short-version items, while 13% and 21% of its objectives received hits on SAT Form A and 17% received hits on Form F.

For alignment to the SAT framework, the SAT items had hits to both "Sentence Completion" and "Passage-Based Reading," and the NAEP items had hits for "Passage-Based Reading" only. The range of knowledge threshold criterion for "Passage-Based Reading" was met for both assessments, with 100% of objectives in "Passage-Based Reading" receiving hits from SAT and NAEP. Regarding "Sentence Completion," it is important to note that, with only one objective in this standard, range of knowledge can be met with one mean hit. Therefore, range of knowledge is not applicable when referring to "Sentence Completion."

Balance of Representation

The NAEP short-version items met or weakly met the balance of representation criteria for all three standards in the NAEP framework. The criterion for balance of representation was weakly met for "Integrate/Interpret," and Panel 2 found that balance was also weak for "Locate/Recall." For the SAT items, for "Integrate/Interpret," in each form, the results across panels are on either side of the threshold between "weakly met" and "not met" for balance of representation. SAT items had very few alignments to the NAEP standards. The two panels differed on the results for "Locate/Recall." For "Critique/Evaluate," both panels found that balance of representation was met on both forms. However, the low categorical concurrence and range of knowledge findings for "Locate/Recall" and "Critique/Evaluate" should be considered in interpreting these balance of representation findings.

In relation to the SAT framework, the SAT short-version items had a balance of representation for both standards (1.0 for "Sentence Completion" and 0.77 for "Passage-Based Reading"). The NAEP items had a weak balance of representation for "Passage-Based Reading" (0.67 and 0.68) only. Regarding "Sentence Completion," it is important to remember that, with only one objective in this standard, balance of representation can be met with one mean hit. Therefore, balance of representation is not applicable when referring to "Sentence Completion."

Overall Conclusions

The following conclusions regarding the alignment of the 2009 NAEP Grade 12 Reading and the SAT Critical Reading 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 SAT?

The greatest commonality between the two tests is their shared emphasis on the broad skills of integrating and interpreting both informational and literary texts. This is evident in the majority of items from both tests aligned to NAEP Standard 2, "Integrate/Interpret," including many to Goal 2.1, "Make complex inferences within and across *both literary and informational texts*."

Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), there is also considerable overlap at the level of more specific skills.

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

Both tests had many of their item alignments to the same NAEP "Integrate/Interpret" objectives, often with similar percentages of alignments. Although there were some differences in emphasis, both tests also had notable percentages of alignments to SAT Objectives B.1.1–B.1.3 and B.1.5. Skills with overlap include inferring/analyzing the following:

- the "main idea" and "author's purpose" (SAT Objective B.1.1 and NAEP Objectives 2.3.a and 2.1.f);
- the "tone and attitude" of an author or character (NAEP Objectives 2.2.a and 2.2.c and SAT Objective B.1.4);
- the use of "rhetorical strategies" (NAEP Objective 2.1.d and SAT Objective B.1.2); and
- connections between ideas, perspectives, or problems (NAEP Objective 2.1.b and SAT Objectives B.1.3 and B.1.5).

Additionally, in the area of greatest content overlap—items on both tests aligned to objectives for NAEP "Integrate/Interpret" and aligned to SAT "Passage-Based Reading" Objectives B.1.1–B.1.5—both tests met the typical threshold criteria for depth of knowledge consistency; that is, most of the items were coded at or above the DOK level of the objectives to which they were aligned.

Despite these similarities, there are some notable differences in emphasis between the two assessments. Both tests assess vocabulary skills. However, NAEP addresses vocabulary exclusively in the context of passage comprehension, while the majority of SAT vocabulary items are in a sentence-completion format, in which context plays a more limited role. This difference reflects NAEP's emphasis on the understanding of word meaning in context; the assessment is not intended to measure students' prior knowledge of word definitions. The SAT sentence-completion items provide some context within the single sentence text, but in many cases, students' success on the items almost certainly depends on their prior knowledge of word definitions.

In addition, panelists found considerably less emphasis in SAT than in NAEP on literal comprehension and critical evaluation, particularly the evaluation of the quality or effectiveness of an author's writing, skills covered in the NAEP standards "Locate/Recall" (locating/recalling specific details and features of texts) and "Critique/Evaluate" (evaluating texts from a critical perspective), respectively. This difference suggests a greater emphasis on these skills in NAEP.

Even with the minimal coverage of NAEP "Locate/Recall" and "Critique/Evaluate" standards by SAT items, all NAEP items found a match in the SAT framework. However, the broad language of the SAT framework can encompass the range of the NAEP items. For example, SAT Goal B.2, "Literal Comprehension," refers to items that "ask what is being said" in a "small but significant portion of a reading passage," a description that can easily accommodate most NAEP

"Locate/Recall" items and objectives. In fact, nearly all items on the NAEP short version that were coded to "Locate/Recall" objectives in the NAEP framework were matched to SAT Goal B.2 in the SAT framework.

Similarly, SAT Objective B.1.3, to which approximately one-quarter of NAEP items aligned, includes "Evaluation," the primary focus of NAEP "Critique/Evaluate." The description in SAT Objective B.1.3 of items that "ask the test taker to evaluate ideas or assumptions in a passage" is compatible at a very general level with NAEP "Critique/Evaluate" objectives addressing the author's point of view, logic, or use of evidence. SAT Objective B.1.2, "Rhetorical Strategies," is also broad enough in its language to make it a reasonable match for some NAEP "Critique/Evaluate" items focused on "author's craft" or use of "literary devices." In the NAEP short version, all items that aligned to "Critique/Evaluate" objectives in the NAEP framework were aligned to either SAT Objectives B.1.2 or B.1.3, or both.

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

With regard to differences in content as described in the NAEP framework, SAT items had limited coverage of the knowledge and skills described by the NAEP standards "Locate/Recall" and "Critique/Evaluate." This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.

With regard to differences in complexity, NAEP items and objectives had a range of depth of knowledge including items at DOK Levels 1, 2, and 3, while SAT items and objectives were coded primarily at Levels 2 and 3.

Overall, the alignment results across the two sets of items and frameworks show a strong area of overlap in their coverage of SAT "Passage-Based Reading" objectives and NAEP "Integrate/Interpret" objectives, as well as some important differences.

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 SAT, 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–SAT reading study, WestEd received from the Governing Board and the College Board very different framework documents with different levels of specificity of content and granularity for NAEP and SAT. In interpreting the study results, it is important to consider that panelists had only eight objectives (including SAT Standard A and Goals B.2 and B.3) to code to in the SAT 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. The larger number of more specific NAEP objectives in comparison with broader SAT objectives increased the likelihood that some objectives would not be matched to items. On the other hand, panelists had more specific information about content in the NAEP objectives to inform their interpretations of the intent of that framework than they had in the SAT objectives.

As discussed in Sections III and V of this report, alignment across the NAEP and SAT 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.

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 SAT 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 12^{th} 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 12^{th} 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 SAT (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 reduced item pools for conducting the within-framework alignments (i.e., NAEP-to-NAEP and SAT-to-SAT). 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–SAT and NAEP–ACCUPLACER 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). *SAT® Skills Insight™ critical reading real SAT questions and answers*. New York, NY: Author.

The College Board. (2008). *SAT® Skills Insight™*. New York, NY: Author.

The College Board. (2010). *The SAT®*. Retrieved October 5, 2010, from http://professionals.collegeboard.com/testing/sat-reasoning

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

Pitoniak, M. J., Reese, C., & Tannenbaum, R. J. (2008a, April). *Technical report on the SAT/NAEP grade 12 preliminary comparability study—mathematics.* Submitted by the Educational Testing Service to the College Board.

Pitoniak, M. J., Reese, C., & Tannenbaum, R. J. (2008b, April). *Technical report on the SAT/NAEP grade 12 preliminary comparability study—reading.* Submitted by the Educational Testing Service to the College Board.

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 ACCUPLACER reading comprehension* (Unpublished report submitted to the National Assessment Governing Board, Contract no. ED-NAG-09-C-001).