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 Mathematics and SAT Mathematics

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

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:
Michael Brown
Linda McQuillen

Study Panelists:
[redacted]

Technical Advisor:
Norman Webb

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

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

Important Notice

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

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

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

Comprehensive Report:
Alignment of 2009 NAEP Grade 12 Mathematics and SAT Mathematics

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 mathematics.

While a typical alignment study explores the alignment between an assessment and a set of standards, this study investigated the degree of alignment between two assessments, assessments that were developed from different frameworks for different purposes. To accomplish its alignment objectives, the Governing Board proposed the use of a bi-directional, multifaceted study design developed by Dr. Norman Webb. This design, as implemented in this current study, comprised a qualitative comparison of the NAEP mathematics framework and the SAT mathematics 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 mathematics framework, 2) the SAT assessment and the SAT mathematics framework, 3) the grade 12 NAEP and the SAT mathematics framework, and 4) the SAT assessment and the NAEP mathematics framework. This bi-directional design allowed for a baseline of alignment to be determined between each assessment and its own framework, which was important in interpreting the degree of cross-framework 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 (i.e., 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 (and range of depth of knowledge), range of knowledge, and balance of representation.

This report addresses the following specific questions:

- What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by 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 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 mathematics subdomains are missing or not aligned?

Summary of Findings

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

NAEP Framework Standards

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

SAT Framework Standards

A. "Number and operations"
B. "Algebra and functions"
C. "Geometry and measurement"
D. "Data analysis, statistics, and probability"

NAEP Assessment to NAEP Framework Alignment

All NAEP items were determined to be codable to the NAEP framework. The NAEP short-version items (42 items) were found to assess all of the five NAEP standards. The largest percentage of NAEP items, approximately one-third, was found to assess "Algebra," while the smallest percentage was found to assess "Data analysis, statistics, and probability."

SAT Assessment to NAEP Framework Alignment

One item in each of the two SAT forms was found to not align to any NAEP objectives, while the remaining 106 SAT items (53 from each form) were found to be codable. Across both forms (54 items each), the SAT items were found to assess all of the five NAEP standards. The largest percentage of SAT items on both forms was aligned to "Algebra," followed by "Number properties and operations" and "Geometry." The smallest percentage of SAT Form D items was aligned to "Data analysis, statistics, and probability," while the smallest percentage of SAT Form E items was aligned to "Measurement."

SAT Assessment to SAT Framework Alignment

The SAT short-version items (40 items) were determined to be codable to the SAT framework, with all items determined to be codable to this framework. Approximately one-third of items from each form were found to align to each of the “Algebra and functions” and “Geometry and measurement” standards. On both SAT forms, the smallest percentage of items was found to assess “Data analysis, statistics, and probability.”

NAEP Assessment to SAT Framework Alignment

NAEP items were found to assess all of the four SAT standards. The largest percentage of NAEP items, nearly one-third, was found to align to “Algebra and functions.” The smallest percentages of items, approximately one-fifth, were aligned to “Number and operations” and “Data analysis, statistics, and probability.” Although nearly all of the 164 NAEP items were determined to be codable to the SAT framework, one panel found three items to not align, and the other panel found two items to not align.

Categorical Concurrence

Categorical concurrence is met for a standard if at least six items are aligned to that standard. For alignment to the NAEP framework, the NAEP items used in the short-version (42 items) were found to meet the typical WAT threshold value of at least six items for categorical concurrence for four of the five standards. In “Data analysis, statistics, and probability,” this criterion was not met. The 108 SAT items were found to meet the criterion for three of the five NAEP standards for Form D and for four of the five NAEP standards for Form E. In Form D, the criterion was on the borderline of being met for “Measurement” and was not met for “Data analysis, statistics, and probability.” In Form E, the criterion was not met for “Measurement.”

For alignment to the SAT framework, both the NAEP items (164 items) and the SAT short-version items (40 items) were found to meet categorical concurrence for all five standards.

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

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

Depth-of-knowledge consistency for a standard is met if at least 50% of the items aligned to an objective in that standard are at or above the DOK level assigned to that objective. For alignment to the NAEP framework, the 42 short-version NAEP items were found to meet depth-of-knowledge consistency in all standards except for “Data analysis, statistics, and probability” (by one panel). The full pool of 108 SAT items also met depth-of-knowledge consistency for all NAEP standards except for “Measurement” for Form E (by one panel).

For alignment to the SAT framework, DOK was analyzed as range of depth of knowledge. NAEP items (164 items) that aligned to the SAT framework were coded at DOK Levels 1, 2, or 3. Most of the NAEP items were coded at DOK Level 2, although a substantial number were coded at DOK Level 1. Similarly, the 40 short-version SAT items were coded at DOK Levels 1, 2, or 3. Most of the SAT short-version items were coded at DOK Level 2, with a substantial number coded at DOK Level 1.

Range-of-Knowledge Correspondence

Range-of-knowledge correspondence is met for a standard if 50% or more of the objectives in that standard have items aligned to them. For alignment to the NAEP framework, the NAEP short-version set of items (42 items) did not meet the criteria for range of knowledge (50% or more of objectives hit) for any standard. No NAEP standard had more than 42% of its objectives hit by items in the short-version, and “Data analysis, statistics, and probability” had the most restricted range of knowledge. This result likely reflects the large number of objectives (130 objectives) relative to the number of items in the short form (42 items) used in this study. For alignment of the SAT items (108 items) to the NAEP framework, range of knowledge was weakly met for “Number properties and operations” and was not met for the other four standards.

For alignment to the SAT framework, the NAEP items (164 items) had a range of knowledge above the 50% criterion for all four standards. For the SAT items (40 items), the range of knowledge criterion was met for all four of the SAT standards.

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.

Both the NAEP short-version items (42 items) and the SAT items (108 items) met the criteria for balance of representation for all five standards in the NAEP framework.

In relation to the SAT framework, the 164 NAEP items met the criteria for balance of representation for “Geometry and measurement” and “Data analysis, statistics, and probability” but only weakly met the criteria for “Number and operations” and “Algebra and functions.” The 40 SAT short-version items met the criteria for balance of representation for all four SAT standards.

Overall Conclusions

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

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

At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining geometry and measurement into one standard. Each framework contains both general and specific objectives, although the SAT objectives, which are presented as content topics without indication of the cognitive level at which that content would be assessed, may be interpreted as more general than the NAEP objectives.

Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are addressed in somewhat similar proportions.

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

The greatest commonality between the two tests is their emphasis at the standard level. This is evident in the distribution of percentages of total hits from both assessments matched to each set of standards. Although there are some differences of emphasis, such as the full NAEP item pool's greater proportion of alignment to SAT "Data analysis, statistics, and probability," and the SAT short-version's greater proportion of alignment to SAT "Geometry and measurement," the proportions of alignments to "Algebra and functions" and "Number and operations" are comparable. There is also considerable overlap among some specific skills, with both assessments addressing many of the same NAEP "Number properties and operations" objectives and SAT objectives (such as N.6, A.10, D.2, and G.1). Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), it is clear that both assessments emphasize a number of the same or closely related skills. These include properties, equivalence, and operations on rational numbers (included in NAEP Goals 1.1 and 1.3 and included in SAT Objective N.2) and properties of two-dimensional shapes (included in NAEP Goals 3.1 and 3.3 and included in SAT Objective G.6).

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 mathematics subdomains are missing or not aligned?

While there is considerable overlap between the two assessments, primarily in the intersection of the NAEP "Algebra" and SAT "Algebra and functions" standards, there are notable differences as well. The SAT items had a somewhat limited range of coverage of the NAEP standards "Measurement," "Geometry," and "Data analysis, statistics, and probability," with several goals receiving few item alignments. Even given the minimal coverage of some of the goals within each NAEP standard by SAT items, however, almost all NAEP items found a match in the SAT framework. The language of the objectives in the SAT framework is sufficiently broad to encompass the range of the NAEP items. For example, SAT Objective A.10, "Basic concepts of algebraic functions," may accommodate most of the items aligning to the seven objectives within NAEP Goal 5.1, "Patterns, relations, and functions." Finally, some NAEP items were found to be uncodable to the SAT objectives. These items assessed skills not present in the SAT framework.

The two tests are also similar in the average DOK levels of items. However, while most items in both tests were found to be at DOK Level 2, NAEP items had a wider range of DOK than did SAT items, with more NAEP items coded to Levels 1 and 3. The Level 3 NAEP items often involved application of concepts through short or extended constructed-response items. Both tests also met depth-of-knowledge consistency overall (with each not meeting this criterion for only one standard as rated by one panel).

Overall, despite differences in alignment at the detailed specific objective level, differences in emphasis at the standard level, and a small difference in ranges of depth of knowledge, there is considerable overlap of content and complexity between the two assessments.

I. Introduction

Purpose

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

The Technical Panel's recommended multi-method approach (National Assessment Governing Board, 2009c) includes conducting content alignment studies; exploring statistical relationships with assessments and outcomes data in postsecondary education and civilian and military job training programs; conducting criterion-based judgmental standard setting activities; and administering national surveys of postsecondary educational institutions. As part of this multi-method approach, the Governing Board contracted WestEd to independently evaluate and report "the extent to which the grade 12 NAEP is aligned in content and complexity to the SAT and to the ACCUPLACER for the two assessments in reading and mathematics" (National Assessment Governing Board, 2009a, p. 3).[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 National Assessment of Educational Progress Grade 12 Mathematics (NAEP) and the SAT Mathematics assessment. Alignment findings from the studies of SAT Critical Reading, ACCUPLACER Reading Comprehension, and ACCUPLACER Mathematics Core Tests are presented in separate reports (WestEd, 2010a, 2010b, 2010c).

Governing Board's Approach to Preparedness

The Governing Board is focusing its conceptualization of 12th grade preparedness on academic qualifications and does not propose to address a range of behavioral and attitudinal aspects of student performance in postsecondary activities that are not measured by NAEP (e.g., time

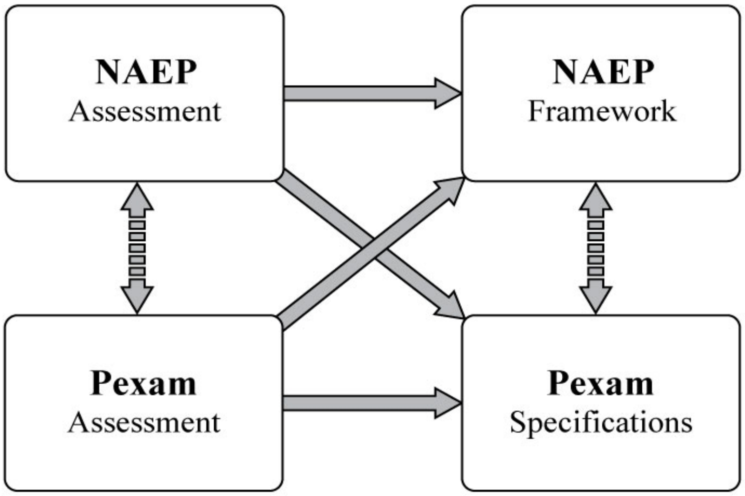
[1] Preliminary comparability studies were conducted by the Educational Testing Service for use by the National Assessment Governing Board and the College Board to determine the feasibility of relating NAEP and SAT in mathematics and reading and of examining alignment of the two more fully (Pitoniak,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

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. (The full text of the resulting study design document is provided in Appendix A.) This study design comprises both a qualitative comparison of the NAEP mathematics framework with the SAT mathematics specifications and a series of alignment activities to investigate the degree of alignment between the pairs of assessments and frameworks/specifications. The qualitative comparisons of each set of frameworks (comparative analyses) are used to inform expectations for alignment, raise potential alignment issues prior to item coding, and inform interpretations of the alignment results. This design is intended to ascertain the degree of alignment of two assessments by comparing how the items on the two assessments represent their respective content domains (National Assessment Governing Board, 2009b, p. 5).

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



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

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 mathematics skills of students at similar ages and stages of academic progress, they serve different purposes for different audiences. NAEP, commonly referred to as "the Nation's Report Card," is administered to representative samples of students across the country, and results are provided at the national level for grade 12. NAEP does not provide results for individual students. 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 (The 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.

Alignment Study

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?

The NAEP–SAT mathematics alignment study discussed in this report was conducted using the Governing Board's study design document developed for grade 12 NAEP alignment studies (National Assessment Governing Board, 2009b). The comparative analysis of the NAEP framework and SAT 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 mathematics framework, 2) the SAT assessment and the SAT mathematics specifications, 3) the grade 12 NAEP and the SAT mathematics specifications, and 4) the SAT assessment and the NAEP mathematics framework. This bi-directional design allowed for a baseline of alignment to be determined between each assessment and its own framework/specifications, which could be used in interpreting the degree of cross-framework/specifications alignment. Alignment criteria used and reported on in this study included categorical concurrence, depth-of-knowledge consistency, range of knowledge, and balance of representation.

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

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

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

This study implemented the study design document developed by Dr. Webb for the Governing Board (National Assessment Governing Board, 2009b) to guide grade 12 NAEP alignment studies in evaluating the degree to which the grade 12 NAEP mathematics assessment aligns in content and complexity to the SAT mathematics 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 framework, 2) the SAT assessment and the SAT framework, 3) the grade 12 NAEP and the SAT framework, and 4) the SAT assessment and the NAEP framework. This bi-directional design allowed for a baseline of alignment to be determined between each assessment and its own framework (within-framework) as well as between each assessment and the other assessment's framework (cross-framework). This within-framework baseline alignment was important in interpreting the degree of cross-framework alignment.

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

The methodology also called for each objective within a standard to be assigned a DOK level. However, the pre-study review of the frameworks indicated that a modification to the study process was required for the SAT mathematics framework. The SAT mathematics framework is organized as a list of topics and does not provide sufficient information to determine the cognitive level at which the knowledge and skills in the objectives would be assessed. Without this information, it would have been impossible for panelists to accurately code the objectives' DOK levels. Therefore, the SAT framework was not coded for DOK. This step was replaced with a review of the SAT objectives, during which panelists carefully reviewed the objectives to gain a level of familiarity with the framework approximating what they would have gained through DOK coding.

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

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

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

Standards and Representation of the Mathematics Content Domain

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

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

The SAT framework represents four categories of assessed content: number and operations questions; algebra and functions questions; geometry and measurement questions; and data analysis, statistics, and probability questions. Within each content category is a bulleted list of more detailed specifications. This framework was determined to lack sufficient information on the intended level of application of skill for panelists to be able to code the objectives for depth of knowledge. As a result, it was determined that panelists could not code the SAT framework for depth of knowledge. WestEd added alphanumeric coding to the framework corresponding to standard (e.g., N) and objective (e.g., N.1) levels. 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 Mathematics Assessment and the SAT Mathematics Assessment

	NAEP Grade 12 Math Assessment	**SAT Math Assessment**
Percentage Distribution of Items by Content Area	Each NAEP mathematics item is developed to measure one of the objectives, which are organized into the four major content areas of mathematics: • Number properties and operations (10%) • Measurement and Geometry (30%) • Data analysis, statistics, and probability (25%) • Algebra (35%)	Each SAT mathematics test is organized into four major content areas: • Number and operations (11–13 items / 20–25%) • Algebra and functions (19–21 items / 35–40%) • Geometry and measurement (14–16 items / 25–30%) • Data analysis, statistics, and probability (6–7 items / 10–15%)
Mathematical Complexity of Items	NAEP test takers spend the following percent of their testing time at each level of complexity: • Low (25%) • Moderate (50%) • High (25%)	SAT items are classified for cognitive complexity as: • Routine • Comprehension • Nonroutine/Insightful
Number of Items	The NAEP pool has 164 total mathematics items. No single student will complete all 164 items. Rather, each student completes two fixed item sets consisting of items from the larger pool.	The SAT mathematics test consists of 54 items: • 44 multiple-choice items • 10 student-produced responses
Item Types	***Multiple choice*** • 4 answer options: 1 correct, 3 incorrect ***Short constructed response*** • 1- or 2-sentence response ***Extended constructed response*** • 1- or 2-paragraph response	***Multiple choice*** • 5 answer options: 1 correct, 4 incorrect ***Student-produced responses*** • Students bubble answers into a grid
Time per Item Type	The intended distribution of items for students is expressed as the percentage of time spent on each item type. • 50% multiple choice • 50% short and extended constructed response	Ten student-produced responses appear in one of the 25-minute sections. The remaining multiple-choice items are distributed across the timed sections as described in the cell below.
Assessment Time	Each student will spend approximately 50 minutes (2 blocks at 25 minutes each) taking the NAEP assessment.	Each student has 70 minutes (two 25-minute sections and one 20-minute section) to take the SAT mathematics assessment.
When Given	NAEP assesses and reports math results every four years.	SAT is offered seven times a year in the U.S. and six times at international sites.

	NAEP Grade 12 Math Assessment	SAT Math Assessment
Testing Population	NAEP is administered to: • 8,000–10,000 students in 200–300 schools • “random samples of students designed to be representative of the nation, different regions of the country, (and) participating states” • 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	SAT is administered to high school students planning to attend college or university.
Accommodations	NAEP allows accommodations specified in an IEP that are routinely used in testing, such as: • Large-print material • Additional time • 1-on-1 or small-group testing • Having directions read • Preferential seating • Breaks during testing • Familiar person testing • Signing of directions • Signing of test items • Magnifying equipment • Template for response • Large marking pen or special writing tool for response • Pointing to answers or responding orally to transcribe Accommodations are offered in combination as needed; for example, students who receive one-on-one testing generally also use extended time. NAEP does **not allow** having items read aloud. For a complete list of NAEP math accommodations see: http://nces.ed.gov/nationsreportcard/about/inclusion.asp#accom_table	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 mathematics accommodations, see: http://www.collegeboard.com

	NAEP Grade 12 Math Assessment	SAT Math Assessment
Calculator Use	The assessment contains blocks (sets of items) for which calculators are not allowed, and calculator blocks, which contain some items that would be difficult to solve without a calculator Two-thirds of the blocks measure students' mathematical knowledge and skills without access to a calculator. One-third of the blocks allow use of a calculator Students allowed to bring any calculator they are accustomed to using in the classroom with some restrictions for test security purposes Scientific calculators supplied to students who do not bring a calculator to use on the assessment	Calculators are permitted for all mathematics questions on the SAT. Every question can be solved without a calculator; however, using a calculator on some questions may be helpful to students. A scientific or graphing calculator is recommended.
Item Scoring	The items are scored as: • Multiple choice: • Incorrect 0 • Correct 1 • Short constructed response: • Incorrect 0 • Partial 1 • Correct 2 • Extended constructed response: • Incorrect 0 • Partial 1 • Essential 2 • Extensive 3 All constructed-response items will be scored using rubrics unique to each item. General principles that apply to these rubrics follow: • Rubrics define minimal, partial, satisfactory, and extended responses. • Students do not receive credit for incorrect responses. • Student responses are coded to distinguish between blank items and items answered incorrectly. • As part of the item review, the testing contractor will ensure a match between each item and the accompanying scoring guide.	The items are scored as: • Multiple choice: • Incorrect ¼ point is subtracted • Correct 1 • Student produced responses: • Incorrect 0 • Correct 1 No points are subtracted for omitted questions.

	NAEP Grade 12 Math Assessment	SAT Math Assessment
Test Scores	**Scaled scores:** Range of 0–300; average scores for groups **Achievement levels:** The numeric scale score range is divided into the following three achievement levels: • **Basic** — This level denotes partial mastery of prerequisite skills and knowledge necessary for proficient work at each grade. • **Proficient** — This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter. • **Advanced** — This level signifies superior performance. Test scores and achievement levels are used to report on the performance of grade 12 students nationally. In 2009, 11 states participated in the first pilot for reporting state NAEP results at grade 12.	**Scaled scores:** Range of 200–800 A raw score is calculated: • Total points for multiple-choice items answered incorrectly 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

Item Pool Selection and Assessment Design

Selection of Item Pools for Alignment Workshop

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

The SAT mathematics assessment is a fixed-form test comprising 54 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 D and E) consisting of 108 unique items (44 multiple-choice and 10 student-produced response items per form, for a total of 54 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-framework analysis was considered sufficient to meet the needs of the study (producing baseline alignment data and providing panelists exposure to each test's items in relation to its own framework) and preferable to not completing the study or having to reconvene panels at a later date. Therefore, with agreement by the study's technical advisor and author of the design document, WestEd and the Governing Board decided to limit the item pools as follows:

NAEP-to-NAEP Alignment

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

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

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

NAEP-to-SAT Alignment

The entire NAEP item pool 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 40 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 40 items to be representative of the range of items in both Form D and Form E. The resulting sample item pool was a reasonable approximation of a representative sample, balancing the number of items with the following characteristics:

- content category;
- item type (multiple-choice versus student-produced response items);
- complexity; and
- set status.

Since the item booklet for Form D was used in both the SAT–SAT and SAT–NAEP studies, the items in Form D were reordered and numbered sequentially so that the 40 items selected for the short-version sample appeared first, followed by the remaining 14 items. The items in Form E were numbered in their original order and contained in a separate booklet used in the SAT–NAEP study.

SAT-to-NAEP Alignment
All 108 items from the two SAT forms (D and E) 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 short-version of 42 items 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.

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

Range-of-Knowledge Correspondence

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

Balance of Representation

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

NAEP and 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 framework; they are described as follows:

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

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

Due to the focus in the Level 4 definition on higher-order thinking tasks carried out over an extended time period, panelists were trained that Level 4 could only apply to tasks (objectives or items) in which both higher-order thinking and extended time were factors, effectively excluding DOK Level 4 as an option for either NAEP or 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 mathematics panels were required to reach joint agreement on the DOK levels of each assessment framework's objectives.[5] As indicated earlier, the SAT objectives were not coded for DOK; therefore, adjudication of DOK of objectives occurred only in relation to the NAEP framework. Prior to alignment coding of the NAEP items, each panel independently coded the NAEP framework for DOK. Once coding was complete, the two panels individually adjudicated to achieve within-panel agreement on DOK levels; the facilitators then met separately to identify and adjudicate differences between the two groups to achieve cross-panel agreement on DOK levels. Upon reaching cross-panel agreement, the facilitators communicated these values to their panelists and entered NAEP framework objectives' DOK values into the WAT. In addition to providing important study data, the DOK adjudication process served a training and calibration purpose, ensuring that panelists were interpreting DOK consistently. Prior to alignment coding of SAT items, each panel independently reviewed SAT objectives to gain familiarity with them. As the system used for data entry and analysis required a DOK value to be entered for each objective, all SAT objectives were assigned a default DOK level of 2.

[5] As stated in the design document, "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).

Adjudication of DOK of Items and Alignment of Items to Frameworks

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

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

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

Alignment Procedure Implemented in the Study

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

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

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, and 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 two mathematics facilitators provided more detailed training in assigning DOK values to objectives to the combined mathematics panels. This initial training included group discussion of mathematics DOK levels and both group and individual practice coding sample objectives drawn from the *WAT Training Manual* (Webb, 2005). When the facilitators determined that the panelists were sufficiently calibrated in their understanding of DOK to begin assigning codes to the frameworks, the panelists separated into their individual panel rooms to register in the WAT. At the end of the first day of the alignment workshop, panelists were given the opportunity to indicate their levels of satisfaction with the training process via an online 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 below (see the study's design document, provided in Appendix A, for a detailed description of each, and see Appendix D for the schedule by which these tasks were conducted). Throughout the week, prior to beginning a new task or after an extended break, facilitators took a few moments to remind panelists of the criteria and tasks at hand.

Review NAEP Framework and Assign DOK Levels to Each Objective

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

Map NAEP Items to the NAEP Framework

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

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

Panelists took a break after discussing and possibly changing their codes, during which time facilitators and project staff began preparing for cross-panel adjudication (the process of ensuring in real time that the panels were functioning as replicate panels). The first steps of this process were for WestEd staff to run the WAT overall results report and prepare the cross-panel adjudication workbook for review and discussion. The facilitators then met separately with WestEd project staff and, usually, the COR, to compare the results of the two groups for discrepancies as outlined in the design document. The facilitators and WestEd project staff reviewed the four alignment criteria: categorical concurrence (reviewing average numbers of items assigned to each objective), depth-of-knowledge consistency (reviewing average percentages of items at, below, and above the DOK level of the assigned objective), range-of-knowledge correspondence (reviewing the percentages of objectives with at least one aligned item), and balance of representation (reviewing index values). Per the design document, discrepancies of greater than five mean hits (categorical concurrence) or five percentage points (depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation), as well as balance of representation index values lower than .7, were investigated

[7] The project director collaborated with the two mathematics facilitators to select a representative range of sample items from the bank of released NAEP items (National Center for Educational Statistics, 2009) and the bank of released SAT items (College Board, 2007). The facilitators then independently coded and reached consensus on DOK levels and alignment to objectives for each item prior to the commencement of this study.

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

to determine whether the differences between panels were systematic or random. As directed by the design document, the facilitators first attempted to resolve areas of discrepancies by discussing observations and panelist opinions raised during the coding process that might have been related to the difference in results. Next, facilitators used the WAT reports to identify specific items that were coded differently by each panel, keeping in mind that panel results are an average across all eight panelists. When relevant items were identified, the facilitators discussed the items and determined whether the difference in coding was reasonable (i.e., not an error), and whether it was random or the result of a systematic difference in interpretation. Facilitators then reported back the outcomes of the cross-panel adjudication (i.e., areas of discrepancy, if any, and whether those discrepancies were systematic or random) to their respective panels, including raising specific items for discussion if necessary. Then, panelists were given the opportunity to change alignment codes if necessary to reflect any changes in their coding judgments. WestEd staff used these final alignment codes in the analysis. Areas of adjudication are discussed in the sub-study results (Section III of this report).

Review the SAT Framework

The design document developed to guide this project’s pilot and operational studies calls for all coding to the NAEP framework to be completed before assigning DOK levels to 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 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. Since the SAT framework did not contain sufficient detail to be coded for DOK, the next step in this alignment workshop’s alignment process was for the panels to review the SAT objectives to become familiar with the content prior to coding. The facilitators met with the combined mathematics panels to discuss the SAT objectives, as well as preliminary decision rules related to the coding of the objectives. In order to proceed with the study using the WAT system, it was necessary to enter a DOK value for each objective. Therefore, facilitators entered “2” as the default value, although it was understood by all that this was not an actual DOK rating for the objectives.

Map SAT Items to the SAT Framework

As with the mapping of NAEP items to the NAEP framework, a subset of 40 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[9] 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

[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.

between the two panels. Within-panel discussions focused on items coded at more than two DOK levels, items coded at non-adjacent levels, and items for which there was no majority of objective codes. Items were discussed, but consensus was not required. Cross-panel adjudication focused on alignment criteria for which there was a discrepancy between panels of greater than five percentage points. Again, consensus was not required, but any issues were communicated to panelists, who had the option of changing any codes. These final alignment codes were used by WestEd staff to determine if the differences between the two panels were, indeed, random and not the result of systematic differences in the application of the protocol or the framework or misinterpretations of the protocol, framework, or items.

Map NAEP Items to the 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 first 42 NAEP items had been assigned DOK levels when being mapped to the NAEP framework, those assigned DOK values were re-entered into the WAT for this task; thus, for the first 42 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.

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-version set of 40 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 40 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 mathematics panelists worked beyond the planned adjournment time each day to complete their work. The time-intensive schedule was anticipated by WestEd and the COR because of the large number of test items and objectives in both mathematics assessments, information gathered in the pilot study, the experiences of the ACT WorkKeys study, and the time estimates in the study design document. Attempts to mediate this included the use of a sample pool for the within-framework sub-studies, as described earlier. WestEd staff monitored the schedule closely, and, in consultation with the COR and facilitators, adjusted the daily schedule as needed. Schedule adjustments were made based on a number of factors, including the importance of keeping the panels synchronized in the tasks they were completing (one panel was not permitted to move ahead to a new coding task before the other had completed it, in case issues arose during cross-panel adjudication that would impact subsequent tasks). WestEd staff tried to allow sufficient break time before starting new tasks, so that panelists would be refreshed and ready to code. Due to the large number of items and objectives to be coded, however, it was necessary to ask panelists to stay past the scheduled end time and/or start earlier in the morning in order to complete the tasks. Panelists were cooperative and started early and stayed late as needed in order to finish each day's tasks. All alignment tasks were completed by both panels by the end of the alignment workshop.

Decision Rules

During the framework analysis and item review conducted prior to the alignment workshop, facilitators developed a preliminary set of decision rules for use by panelists. Facilitators reviewed the preliminary decision rules with panelists and instructed panelists in their use prior to alignment coding, ensuring that panelists were comfortable with the decision rules. Throughout the alignment coding sessions, panelists were allowed to develop additional decision rules and modify existing decision rules to clarify potential ambiguities in assessments and assessment frameworks, thereby promoting consistency in coding both within and across panels; decision rule 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 Mathematics Framework for Alignment: Decision Rules

1. The objectives within the Algebra standard will be interpreted as aligning primarily to items containing one or more variables and not items containing only numerical expressions.
2. The objectives within the Number Operations and Properties standard will be interpreted as aligning primarily to numerical items; however, consideration is also to be given to items containing one or more variables.
3. The primary intent of objectives containing wording similar to the following is to assess mathematics in (problem-solving situations (in either a real-world or mathematical context), as opposed to the performance of simple procedures or algorithms).[10]
    - 1.1.g, "Represent, interpret, or compare expressions or problem situations involving absolute values."
    - 1.3.f, "Solve application problems involving numbers, including rational and common irrationals."
    - 1.4.c, "Use proportions to solve problems (including rates of change)."
4. Some objectives contain multiple parts separated by the word "and" (see 1.5.f below). The intent of the objective may or may not be to assess all parts. If an item addresses only one part of the objective, panelists are asked to look for an alternative primary code. If an alternative is not available, panelists are to note in the WAT that the item does not assess the entire objective.
    - 1.5.f, "Recognize properties of the number system (whole numbers, integers, rational numbers, real numbers, and complex numbers) and how they are related to each other, and identify examples of each type of number."
5. An objective that addresses expressions may also be aligned with an item containing an equation if symbolic manipulation across the equal sign is not required to answer the question.

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

SAT Mathematics Framework for Alignment: Decision Rules

1. In the SAT specifications, parentheticals are not necessarily exclusive. As an example, G.1, "Points and lines in the plane (locus, parallel, perpendicular) including use of geometric notation (length, segments, lines, rays, and congruence)," may include other types of points and lines in the plane and/or other types of geometric notation.

The following content-specific decision rules were developed as guidelines for coding the NAEP items to the SAT specifications.

2. Items assessing numerical reasoning and numerical proofs are coded to N.7, "Logic/logical reasoning," with a note (e.g., "The item assesses the logic of numbers").
3. Items assessing imaginary numbers in a context involving quadratic equations are coded to A.7, "Quadratics." Other items assessing imaginary numbers should be coded to SAT N, "Numbers and Operations," at the standard level, if numerical, or SAT A, "Algebra and functions," at the standard level, if the item contains variables.
4. Items assessing exponential, logarithmic, and inverse functions are coded to SAT A, "Algebra and functions," at the standard level, with a note (e.g., "The item assesses exponential functions, which are not directly assessed by the SAT framework").
5. Items assessing geometric proofs, networks, polar coordinates, vectors, or scale drawings are coded to SAT G, "Geometry," at the standard level, with a note (e.g., "The item assesses geometric proofs, which are not directly assessed by the SAT framework").
6. Items assessing geometric transformations should be coded to either G.8, "Geometric perception," or G.9, "Coordinate geometry," with a note (e.g., "The item assesses transformations").
7. Trigonometric functions that specifically assess special triangles are coded to G.4, "Special triangles." Other items assessing trigonometric functions are uncodable to the SAT framework.
8. Items assessing knowledge of spreadsheets are uncodable to the SAT framework, because this skill is not included in any objective.
9. Items assessing data analysis and/or statistics that are not codable directly to D.1, "Data interpretation," or D.2, "Statistics," are coded to SAT D, "Data analysis, statistics, and probability," at the standard level, with a note (e.g., "The item assesses the normal distribution, which is not directly assessed by the SAT framework").
10. Items assessing surveys and statistical variability are coded to SAT D, "Data analysis, statistics, and probability," at the standard level, with a note (e.g., "The item assesses standard deviation, which is not directly assessed by the SAT framework").

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.

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

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

Curriculum and Instruction. She has taught for over 40 years, specializing in the areas of mathematics and special education. Ms. McQuillen received a BS in secondary education from Northern State College, and a MS in exceptional education from the University of Wisconsin, Milwaukee.

Panel Criteria for Recruitment and Panelist Qualifications

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

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

In order to ensure that the panelists did not hold biases toward any of the assessments included in the study, panelists with substantial involvement in the development of either NAEP or 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 applicants, two potential panelists reported general exposure to NAEP through involvement in prior NAEP-related alignment and other research projects; this exposure was deemed by the Governing Board COR to not be problematic for the purposes of this alignment workshop.

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

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

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

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

- On the first panel, 63% of panelists (5 of 8) reported experience in both secondary and postsecondary mathematics education; 25% (2 of 8) had secondary teaching experience only; and 13% (1 of 8) had postsecondary teaching experience only.
- On the second panel, 50% of panelists (4 of 8) reported experience in both secondary and postsecondary mathematics education; 25% (2 of 8) had secondary teaching experience only; and 25% (2 of 8) had postsecondary teaching experience only.

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 five women on Panel 1 and four women on Panel 2, as was representation of advanced degrees, with four doctoral degrees represented on each of Panels 1 and 2. Panelists represented a range of geographic areas, including the Northeast (Connecticut, Maryland, Pennsylvania), the South (Georgia, Kentucky, Louisiana), the Midwest (Illinois, Iowa, Indiana, Minnesota), and the West (Arizona, California, Colorado, Hawaii). WestEd was able to achieve some degree of race/ethnicity diversity. Panelists identified themselves as White/Caucasian/of European descent (10); Black/African-American (2); Hispanic or Latino (1); Hispanic or Latino / White/Caucasian/ of European descent (1); and Multi-Racial (1). One panelist did not indicate his/her race/ethnicity. A list of panelists organized by panel follows.

Mathematics Panel 1

[redacted]

[redacted]

[redacted]

[redacted]

Department heads from top-tier national and regional institutions were contacted to solicit referrals and to recruit as potential candidates.

[redacted]

[redacted]

[redacted]

[redacted]

Mathematics Panel 2

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[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 the *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 re-familiarize themselves with 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 as observers 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 facilitator training:

- Alignment workshop objectives and 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 this NAEP–SAT study, the study's lead facilitator developed the comparative analysis to document the similarities and differences between the NAEP and SAT frameworks. 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 preliminary 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 *Mathematics 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 panelists 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 combined mathematics panels received training on assigning DOK levels to objectives, using practice objectives drawn from the *WAT Training Manual* (Webb, 2005). Additional training on assigning DOK levels to items and assigning items to objectives was subsequently provided, using sample items drawn from the *NAEP Sample Questions, Grade 12, 2009* (National Center for Educational Statistics, 2009); and from the *SAT® Skills Insight™ Mathematics Real SAT Questions and Answers* (College Board, 2007). These sample items were selected by the mathematics facilitators to represent a range of item types, DOK levels, and objective alignments, and are included in Appendix E.

On-Site Security of Materials

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

- Shipping 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, the 164 items were organized in block and item order, numbered sequentially within each block, except that the 42 items identified for coding to the NAEP framework were listed first, also in block and item order. Each item was presented on a separate page, with grade, block code, NAEP identification and WestEd sequence numbers, item type, and answer key indicated at the top of the page.

For SAT, the items from the two forms were bound separately. In the Form D booklet, the short-version of 40 items was listed first, followed by the remaining 14 Form D items. The Form E 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 an online survey system (SurveyMonkey).[12]
- An evaluation-of-process questionnaire was administered at or near the end of each of the second, third, and fourth days of the alignment workshop. These questionnaires were used to monitor panelists' understanding of the process, and to solicit questions, concerns, and other feedback from panelists regarding that day's activities. These questionnaires were administered via the online SurveyMonkey system.
- An end-of-study 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

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

and/or needs for additional training; these results are summarized in Section IV of this report. Full responses to the process questionnaire 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 to panelists to express any thoughts, concerns, or questions that remained regarding the assessments, objectives of the overall study, and projected use of study results.

WAT System

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

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

During the workshop, WAT system server errors outside the control of the project were experienced for five of the mathematics WAT studies, making it impossible to download certain WAT reports for these studies. In these cases, duplicate studies were created in the WAT, and the raw data tables or, when the raw data tables were not available, the handwritten codes, were 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.

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

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

Pilot Study: Lessons Learned

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

Sequence of Study Steps

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

Within-Panel Adjudication

- Facilitators may share their own alignment interpretations to foster group discussions and help clarify understandings and interpretations, but care should be taken to ensure 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.
- Identify and provide additional information, if available, to elaborate on the ACCUPLACER frameworks used for coding. [13]

Facilitator Training

- Provide facilitators with assessment frameworks and sample items for review at least two weeks in advance of the study. As facilitators code sample assessment items to the frameworks, they will identify any preliminary decision rules and determine where coding and adjudication discrepancies and areas of potential confusion might exist prior to the study.
- Refine facilitator training to include additional training on the WAT system, tailored specifically for this study, and the use of the WestEd Excel workbook tool as well as the logistics of the methodology.

Panelist Training

- Provide frameworks and other preparatory materials to panelists at least two weeks prior to the study as mandatory reading material for the session.
- Refine panelist training to address and/or emphasize the areas identified in the Pilot Study Report as needing clarification or specifications: alignment criteria, including

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

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 reading only. For SAT mathematics, the two forms were analyzed separately. Because there was no overlap of items across the forms, 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 DOK of the NAEP framework and NAEP and SAT assessment items are discussed. Finally, the results of the four sub-studies are presented.

Reliability and Interrater Agreement

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

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

Pairwise comparison statistics are also calculated to show the interrater agreement for panelists’ judgments of alignment of items to objectives in the frameworks. Interrater reliability of these judgments is reported at the more specific objective level (i.e., the degree to which panelists reached the same judgment of the objective[s] tested by an item) and the more general standard level (i.e., the degree to which panelists reached the same judgment of the standard containing the objective[s] tested by an item). Objective and standard pairwise comparison are calculated as follows: for each pair of reviewers, “find the reviewer who coded the greater number of objectives to this item, and call this number *n*. Now take the number of entries the two reviewers agree on and divide this by *n*. This is the *agreement* between the two reviewers. Perform this calculation for all possible pairs of reviewers, and take the sum of the agreements. Then divide this sum by the total number of pairs of reviewers. This is the *pairwise agreement* value for the given assessment item . . . The pairwise agreement for objectives is averaged over all the assessment items to give the *pairwise agreement for objectives* statistic for the alignment study as a whole" (Webb, 2005, p. 115). It is typical that objective pairwise comparison values are lower than those for standard pairwise comparison, because objectives tend to be more specific applications of a broader topic defined in a standard.

Table 2. Interrater Agreement of Panels by Sub-Study

Sub-Study	Panel 1	Panel 2
Sub-Study 1: NAEP to NAEP	*DOK* Intraclass Correlation: 0.9262 Pairwise Comparison: 0.6658 *Objective, Standard* Objective Pairwise Comparison: 0.6693 Standard Pairwise Comparison: 0.9331	*DOK* Intraclass Correlation: 0.9255 Pairwise Comparison: 0.5549 *Objective, Standard* Objective Pairwise Comparison: 0.7047 Standard Pairwise Comparison: 0.8778
Sub-Study 2: SAT to NAEP	*Form D* *DOK* Intraclass Correlation: 0.9026 Pairwise Comparison: 0.7401 *Objective, Standard* Objective Pairwise Comparison: 0.5813 Standard Pairwise Comparison: 0.8133 *Form E* *DOK* Intraclass Correlation: 0.8175 Pairwise Comparison: 0.7209 *Objective, Standard* Objective Pairwise Comparison: 0.6026 Standard Pairwise Comparison: 0.8585	*Form D* *DOK* Intraclass Correlation: 0.8755 Pairwise Comparison: 0.7526 *Objective, Standard* Objective Pairwise Comparison: 0.5473 Standard Pairwise Comparison: 0.7859 *Form E* *DOK* Intraclass Correlation: 0.8121 Pairwise Comparison: 0.6442 *Objective, Standard* Objective Pairwise Comparison: 0.6372 Standard Pairwise Comparison: 0.835
Sub-Study 3: SAT to SAT	*DOK* Intraclass Correlation: 0.9051 Pairwise Comparison: 0.7107 *Objective, Standard* Objective Pairwise Comparison: 0.596 Standard Pairwise Comparison: 0.8056	*DOK* Intraclass Correlation: 0.8728 Pairwise Comparison: 0.7634 *Objective, Standard* Objective Pairwise Comparison: 0.7584 Standard Pairwise Comparison: 0.9098
Sub-Study 4: NAEP to SAT	*DOK* Intraclass Correlation: 0.8923 Pairwise Comparison: 0.6405 *Objective, Standard* Objective Pairwise Comparison: 0.6073 Standard Pairwise Comparison: 0.8594	*DOK* Intraclass Correlation: 0.9035 Pairwise Comparison: 0.7169 *Objective, Standard* Objective Pairwise Comparison: 0.7174 Standard Pairwise Comparison: 0.8788

Looking across panels, Table 2 shows that interrater agreement (within-panel) values for each panel appeared comparable for DOK and for alignment. Interrater agreement for DOK (intraclass correlation and pairwise comparison) met the threshold for "good," as defined by the WAT, for all sub-studies for both panels, with one exception: for Sub-Study 1 (NAEP-to-NAEP), Panel 2 had a DOK pairwise comparison value below the "reasonable" range, with a value of 0.5549. Standard pairwise comparison values were good for all studies for both panels. For match to objective, objective pairwise comparison values met the threshold for "good" or "reasonable" for all sub-studies except for Sub-Study 2 (SAT-to-NAEP) Form D in both Panel 1 and Panel 2 and

Sub-Study 3 (SAT-to-SAT) for Panel 1. For Sub-Study 2, Panels 1 and 2 each had an objective pairwise comparison value below the “reasonable” range and above the “poor” range, with values of 0.5813 and 0.5473, respectively; for Sub-Study 3, Panel 1 had an objective pairwise comparison value just below the “reasonable” range, with a value of 0.596.

Lower objective-level pairwise comparison values can result from overlapping or unclear objectives within or across standards, as well as from items being coded to multiple objectives. Indeed, facilitators and panelists found that there were overlapping objectives in both the NAEP and SAT frameworks and that test items were sufficiently complex to be reasonably aligned to a number of different objectives, often within the same goal. Overall, the interrater agreement levels warrant confidence in the reliability of each panel’s findings and the overall conclusions of the study.

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

DOK Levels of the NAEP Framework

Panelists assigned DOK levels to each objective in the NAEP framework. The within-panel DOK ratings were then compared across panels, and the two facilitators reached consensus on the final DOK ratings for each objective, discussing them with the combined mathematics panels as appropriate. Consensus DOK values for the NAEP framework are shown in Table 3. DOK ratings were assigned to the 130 specific objectives. These ratings are reported at the goal and standard level in the table. DOK ratings for each objective can be found in Appendix C. As explained in Section II, the SAT specifications did not contain sufficient information to be coded for DOK.

Table 3. DOK Findings for the NAEP Mathematics Framework

NAEP Framework	# of Objectives	# and % of Obj. at DOK 1	# and % of Obj. at DOK 2	# and % of Obj. at DOK 3	Average DOK
1.1	4	2 (50%)	2 (50%)	-	1.5
1.2	3	1 (33%)	-	2 (67%)	2.33
1.3	5	3 (60%)	2 (40%)	-	1.4
1.4	2	-	2 (100%)	-	2
1.5	4	2 (50%)	2 (50%)	-	1.5
1.6	2	-	-	2 (100%)	3
1 overall	**20**	**8 (40%)**	**8 (40%)**	**4 (20%)**	**1.8**
2.1	6	-	6 (100%)	-	2
2.2	5	1 (20%)	4 (80%)	-	1.8
2.3	7	1 (14%)	6 (86%)	-	1.86
2 overall	**18**	**2 (11%)**	**16 (89%)**	**-**	**1.89**

NAEP Framework	**# of Objectives**	**# and % of Obj. at DOK 1**	**# and % of Obj. at DOK 2**	**# and % of Obj. at DOK 3**	**Average DOK**
3.1	4	1 (25%)	3 (75%)	-	1.75
3.2	6	1 (17%)	4 (67%)	1 (17%)	2
3.3	7	1 (14%)	6 (86%)	-	1.86
3.4	8	1 (13%)	7 (88%)	-	1.88
3.5	5	-	-	5 (100%)	3
3 overall	**30**	**4 (13%)**	**20 (67%)**	**6 (20%)**	**2.07**
4.1	6	-	5 (83%)	1 (17%)	2.17
4.2	7	-	7 (100%)	-	2
4.3	5	1 (20%)	2 (40%)	2 (40%)	2.2
4.4	9	1 (11%)	8 (89%)	-	1.89
4.5	5	-	2 (40%)	3 (60%)	2.6
4 overall	**32**	**2 (6%)**	**24 (75%)**	**6 (19%)**	**2.13**
5.1	7	-	6 (86%)	1 (14%)	2.14
5.2	7	-	4 (57%)	3 (43%)	2.43
5.3	7	3 (43%)	4 (57%)	-	1.57
5.4	6	3 (50%)	2 (33%)	1 (17%)	1.67
5.5	3	-	1 (33%)	2 (67%)	2.67
5 overall	**30**	**6 (20%)**	**17 (57%)**	**7 (23%)**	**2.03**
ALL	**130**	**22 (17%)**	**85 (65%)**	**23 (18%)**	**2.01**

As shown in Table 3, across all standards and the 130 objectives, the distribution of DOK levels was 17% (22) at Level 1, 65% (85) at Level 2, and 18% (23) at Level 3, for an average DOK of 2.01. For each standard except NAEP Standard 1, “Number properties and operations,” the majority of objectives were assigned DOK Level 2. The 20 objectives in NAEP Standard 1 were distributed between DOK Level 1 (40%), Level 2 (40%), and Level 3 (20%), for an average DOK level of 1.8. The 18 objectives in Standard 2, “Measurement,” were divided between DOK Level 1 (11%) and Level 2 (89%), for an average DOK level of 1.89. The 30 objectives in Standard 3, “Geometry,” were assigned DOK Level 1 (13%), Level 2 (67%), or Level 3 (20%), for an average DOK level of 2.07. The 32 objectives in Standard 4, “Data analysis, statistics, and probability,” were assigned DOK Level 1 (6%), Level 2 (75%), or Level 3 (19%), for an average DOK level of 2.13. The 30 objectives in Standard 5, “Algebra,” were assigned DOK Level 1 (20%), Level 2 (57%), and Level 3 (23%), for an average DOK level of 2.03. The two standards with the highest percentage of Level 3 objectives were “Algebra” and “Geometry.” The standards with the highest average overall DOK were “Data analysis, statistics, and probability” and “Geometry.”

DOK Levels of the SAT Framework

The SAT framework used in this study consists of five standards and 29 objectives. As described earlier, because the objectives in the specifications contain content topics but do not make explicit the intended level of application of knowledge and skills, the SAT specifications were not coded for DOK.

DOK Levels of the Test Items

Panelists assigned each item a DOK rating, independent of any content alignment. Because panelists were not required to reach consensus on the DOK values of items, these ratings were not consensus ratings, and interrater agreement for DOK is addressed in Table 2. The average DOK levels of the NAEP items in the short-form set of 42 items used for the NAEP-to-NAEP study were 1.57 for Panel 1 and 1.67 for Panel 2. The average DOK levels of the NAEP items in the complete set of 164 items used for the NAEP-to-SAT study were 1.62 for Panel 1 and 1.82 for Panel 2. The average DOK levels for the SAT items in the short-form set of 40 items used for the SAT-to-SAT study were 1.76 for Panel 1 and 1.90 for Panel 2. The average DOK levels of the SAT items in the complete set of 108 items (54 per form) used for the SAT-to-NAEP study were 1.79 for Panel 1 and 1.83 for Panel 2 in Form D, and 1.80 for Panel 1 and 1.60 for Panel 2 in Form E. The comparison of the DOK levels of the test items with the DOK levels of the NAEP objectives to which they align is addressed in the depth-of-knowledge consistency analyses for Sub-Studies 1 and 2 later in this section. The range of DOK levels of the test items aligned to the SAT framework is addressed in the analyses for Sub-Studies 3 and 4.

Alignment Results by Sub-Study

The alignment results of each sub-study are presented in the following sections. As discussed in Section II of this report, the order in which the sub-studies were conducted was modified so that each assessment would be coded to its own framework or specifications 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 surveys are provided in Appendices N and O):

- Sub-Study 1—NAEP Items (Short Version) to NAEP Framework
- Sub-Study 2—SAT Items (Forms D and E) 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, reviewers evaluated the alignment between the NAEP items and the NAEP framework. A short-form sample of 42 items was analyzed. The results of Sub-Study 1 are presented in Tables 4–8.

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

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

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

As shown in Table 5, all 42 items were coded to at least one objective.

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

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

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	43.75	100%	43.13	100%
Uncodable	0.00	0%	0.00	0%
Total	43.75		43.13	

For the 42 items, the total mean hits for the two panels were 43.75 and 43.13. These numbers exceed 42 because some items were coded to multiple objectives by one or more panelists. No uncodable ratings were assigned.

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

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

	Panel 1			**Panel 2**		
Standards	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
1–Number properties and operations	9.63	22	22	9.38	22	22
2–Measurement	7.13	16	16	6.75	16	16
3–Geometry	7.13	16	16	7.00	16	16
4–Data analysis, statistics, and probability	4.88	11	11	5.13	12	12
5–Algebra	15.00	34	34	14.88	34	34
Total	43.75	100	100	43.13	100	100

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

All NAEP standards received hits from NAEP items in the short-version subset, with a distribution as indicated in Table 6. Of the five standards, Standard 5, “Algebra,” received the greatest number of mean hits in both panels (15.00 and 14.88 for Panels 1 and 2, respectively), making up 34% of the item set for each panel. Standard 4, “Data analysis, statistics, and probability,” received the fewest mean hits in both panels: 4.88 mean hits in Panel 1 (11% of total hits) and 5.13 mean hits in Panel 2 (12% of total hits). As mentioned in Section II, it should be noted that, although it was considered sufficiently representative across a number of characteristics and small enough to allow for completion of coding, the short-version item sample was overrepresentative of “Number properties and operations” by approximately 9 percentage points (4–5 items) and underrepresentative of “Data analysis, statistics, and probability” by approximately 12 percentage points (5 items).

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

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

			Panel 1		Panel 2	
Standards	**Goals**	**Objectives**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
1	1.1	1.1.d	1	2	0.88	2
		1.1.f	1	2	0.88	2
		1.1.g	0.25	1	0	0
		1.1.i	0	0	0	0
	1.2	1.2.b	1.5	3	0	0
		1.2.c	0	0	0	0
		1.2.d	0	0	0	0
		1.2	0.25	1	0.25	1
	1.3	1.3.a	0	0	0.25	1
		1.3.b	1.5	3	2.5	6
		1.3.c	0.75	2	0.88	2
		1.3.d	0	0	0.13	0
		1.3.f	0	0	1	2
	1.4	1.4.c	0.13	0	0.25	1
		1.4.d	1.13	3	1.25	3
	1.5	1.5.c	1	2	1	2
		1.5.d	0	0	0	0
		1.5.e	0.13	0	0.13	0
		1.5.f	0	0	0	0
	1.6	1.6.a	0	0	0	0
		1.6.b	1	2	0	0
2	2.1	2.1.b	0.13	0	0.13	0
		2.1.c	0	0	1	2
		2.1.d	0	0	0	0
		2.1.f	1.75	4	0.38	1
		2.1.h	0.13	0	0	0
		2.1.i	0.63	1	1.13	3
	2.2	2.2.a	0.75	2	0.75	2
		2.2.b	0.75	2	0.25	1
		2.2.d	0.13	0	0	0
		2.2.e	0.13	0	0	0
		2.2.f	1.75	4	2	5
	2.3	2.3.a	0	0	0	0
		2.3.b	0	0	0.13	0
		2.3.c	1	2	0.88	2
		2.3.d	0	0	0	0
		2.3.e	0	0	0.13	0
		2.3.f	0	0	0	0
		2.3.g	0	0	0	0

			Panel 1		Panel 2	
Standards	**Goals**	**Objectives**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
3	3.1	3.1.c	0	0	0	0
		3.1.d	0.13	0	0	0
		3.1.e	0	0	0.13	0
		3.1.f	0	0	0	0
	3.2	3.2.a	0	0	0.13	0
		3.2.b	0	0	0	0
		3.2.c	0.13	0	0	0
		3.2.d	0.88	2	0.88	2
		3.2.e	0	0	0	0
		3.2.g	0	0	0	0
	3.3	3.3.b	0.25	1	0.38	1
		3.3.c	0	0	0	0
		3.3.d	1.13	3	0.75	2
		3.3.e	0.25	1	0	0
		3.3.f	0.88	2	1	2
		3.3.g	0	0	0	0
		3.3.h	0.63	1	0.88	2
	3.4	3.4.a	1	2	1.13	3
		3.4.b	0	0	0	0
		3.4.c	1	2	1	2
		3.4.d	0	0	0	0
		3.4.e	0	0	0	0
		3.4.f	0	0	0	0
		3.4.g	0	0	0	0
		3.4.h	0.88	2	0.75	2
	3.5	3.5.a	0	0	0	0
		3.5.b	0	0	0	0
		3.5.c	0	0	0	0
		3.5.d	0	0	0	0
		3.5.e	0	0	0	0
4	4.1	4.1.a	1.38	3	1.88	4
		4.1.b	0.13	0	0.13	0
		4.1.c	0	0	0	0
		4.1.d	0.63	1	0.25	1
		4.1.e	0	0	0	0
		4.1.f	0	0	0	0
	4.2	4.2.a	0.13	0	0	0
		4.2.b	0	0	0	0
		4.2.c	0	0	0	0
		4.2.d	0.25	1	0.25	1
		4.2.e	0	0	0	0
		4.2.f	0	0	0	0
		4.2.g	0	0	0	0
	4.3	4.3.a	0	0	0	0

			Panel 1		Panel 2	
Standards	Goals	Objectives	Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		4.3.b	0	0	0	0
		4.3.c	0.5	1	0	0
		4.3.d	0	0	0	0
		4.3.e	0	0	0	0
	4.4	4.4.a	0	0	0	0
		4.4.b	0.13	0	0.13	0
		4.4.c	0.63	1	0.75	2
		4.4.d	0	0	0	0
		4.4.e	0	0	0	0
		4.4.h	0.63	1	0.88	2
		4.4.i	0	0	0	0
		4.4.j	0.13	0	0	0
		4.4.k	0	0	0	0
	4.5	4.5.a	0.38	1	0.88	2
		4.5.b	0	0	0	0
		4.5.c	0	0	0	0
		4.5.d	0	0	0	0
		4.5.e	0	0	0	0
5		5	0	0	0.13	0
	5.1	5.1.a	1	2	1.63	4
		5.1.b	1.88	4	1	2
		5.1.e	1	2	2.13	5
		5.1.g	0	0	0	0
		5.1.h	1.13	3	0	0
		5.1.i	0.88	2	0	0
		5.1.j	0	0	0	0
	5.2	5.2.a	0.75	2	1.88	4
		5.2.b	0.25	1	0.13	0
		5.2.d	0.25	1	0.25	1
		5.2.e	0	0	0	0
		5.2.f	1	2	0.88	2
		5.2.g	0	0	0	0
		5.2.h	0	0	0	0
	5.3	5.3.b	0.13	0	0.25	1
		5.3.c	1.13	3	1	2
		5.3.d	1.13	3	0.88	2
		5.3.e	1.88	4	1.5	3
		5.3.f	0.88	2	0.88	2
		5.3.g	0	0	0	0
		5.3.h	0	0	0	0
	5.4	5.4.a	0.75	2	0.13	0
		5.4.c	0	0	0.13	0
		5.4.d	0	0	0.13	0
		5.4.e	0	0	0	0

			Panel 1		Panel 2	
Standards	Goals	Objectives	Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
		5.4.f	0	0	0	0
		5.4.g	0	0	0	0
	5.5	5.5	0.13	0	0	0
		5.5.a	0	0	2	5
		5.5.b	0.88	2	0	0
		5.5.c	0	0	0	0

As shown in Table 7, 70 of the 130 objectives received one or more hits; additionally, two goals and one standard received at least one hit directly at the goal or standard level. The two objectives with the greatest number of mean hits were:

- 1.3.b, “Perform arithmetic operations with real numbers, including common irrational numbers” (1.5 mean hits in Panel 1 and 2.5 mean hits in Panel 2)
- 2.2.f, “Construct or solve problems involving scale drawings” (1.75 mean hits in Panel 1 and 2.0 mean hits in Panel 2)

The two items coded to Objective 1.3.b are skills-based and involve operations on rational numbers. Though the coding of these two items by the panelists varied somewhat, almost all the codes for these two items fell within Goals 1.3, “Number operations,” or 1.2, “Estimation,” with 50% or more of the hits across both panels to Objective 1.3.b.

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

Within Standard 3, “Geometry,” objectives within Goals 3.3 and 3.4 received the majority of hits; objectives within Goal 3.3, “Relationships between geometric figures,” received 3.125 and 2.0 mean hits in Panels 1 and 2, respectively, and objectives within Goal 3.4, “Position, direction, and coordinate geometry,” received 2.875 mean hits in both Panels 1 and 2. The objectives listed under Goals 3.3 and 3.4 tend to be broad in scope, and items aligned to them can address a range of skills. Across the goals for the other two standards—Standard 1, “Number properties and operations,” and Standard 4, “Data analysis, statistics, and probability”—the coding of items was distributed across a range of objectives, often with fewer than 1.00 mean hits per objective.

As shown in Table 7, of the 130 objectives, 93 in Panel 1 and 98 in Panel 2 received two or fewer hits (0–.25 mean hits). Two of the standards—Standard 3, “Geometry,” and Standard 4, “Data analysis, statistics, and probability”—had the greatest number of objectives receiving two or fewer hits (23 of 30 in Panel 1 and 22 of 30 in Panel 2 for Standard 3, and 26 of 32 in Panel 1 and 28 of 32 in Panel 2 for Standard 4). However, any conclusions drawn from the data must take into account the large number of objectives (130) relative to the number of items (42) in this study.

The comments recorded by the panelists in debrief questionnaires expressed the observation that some important objectives were not addressed in the NAEP short-version sample. For example, panelists commented that there were few items coded to objectives under Goal 1.5, "Properties of number and operations."

When assigning objectives to items, an item can possibly be aligned to more than one objective. For example, one item involved operations on decimal numbers and also assessed estimation and place value. The coding of this item to the objectives in Goal 1.3 by eight panelists across both panels reflects their judgment that the item is primarily about operations, while the coding of this item to the objectives in Goal 1.2 by six panelists across both panels reflects their judgment that the item is primarily about estimation.

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

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

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

	Alignment Criteria							
	Categorical Concurrence (mean hits)		**Depth-of-Knowledge Consistency** (% of hits at or above level of standard)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
1–Number properties and operations	9.62	8.33	72	87	42*	32**	0.91	0.75
2–Measurement	7.12	6	61	76	31**	27**	0.85	0.76
3–Geometry	7.12	6.22	54	59	22**	20**	0.96	0.87
4–Data analysis, statistics, and probability	4.88**	4.56**	46*	56	14**	12**	0.92	0.78
5–Algebra	15	13.22	65	73	39**	30**	0.85	0.73

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

Of the 42 NAEP assessment items analyzed, all (42) were found to match or “hit” objectives. Using the typical WAT threshold value of six mean hits, categorical concurrence was met for Standard 1, “Number properties and operations,” with 9.62 and 8.33 mean hits to the standard. As shown in Table 8, categorical concurrence was also met for Standard 2, “Measurement,” with 7.12 and 6 mean hits to the standard, and Standard 3, “Geometry,” with 7.12 and 6.22 mean hits to the standard. Standard 5, “Algebra,” received the most hits, with 15 and 13.22 mean hits to the standard. Categorical concurrence was not met for Standard 4, “Data analysis, statistics, and probability,” with 4.88 and 4.56 mean hits to the standard. However, it is important to note that there was a distribution of alignment across the standards, and in a larger item pool there would be a greater chance of meeting this numerical threshold.

The depth-of-knowledge consistency criterion was met by both panels for four of the five standards, with more than 50% of the items at or above the DOK level of the standard to which they aligned. The two standards rated lowest in terms of depth-of-knowledge consistency were “Geometry,” with 54% and 59% of hits for items at or above the DOK level of the objective to which they were coded, and “Data analysis, statistics, and probability,” with 46% and 56%. The objectives for these two standards had the highest DOK ratings of the five standards. Discrepancies of greater than five percentage points between panels exist for four of the five standards. However, due to the relatively small numbers of items and mean hits causing the percentage differences at each standard, the discrepancy was not considered atypical.

Using the typical WAT threshold values, the range of knowledge criterion was not met for any standard, with the exception of “Number properties and operations,” for which Panel 1 found the range of knowledge to meet the “weak” threshold value, with 42%. Panelists found that between 12% and 42% of the objectives in each standard were assessed in the item sample. This is largely due to the relatively small number of items being aligned to a large number of objectives. Given the full item pool, it would be expected that a greater range of objectives would receive hits.

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

Sub-Study 2—SAT Items (Forms D and E) to NAEP Framework

In Sub-Study 2, reviewers evaluated the alignment between the SAT items and the NAEP framework. Two 54-item forms (Forms D and E) of the SAT assessment were aligned with the NAEP framework, for a total of 108 items. The results of Sub-Study 2 are presented in Tables 9–13.

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

Table 9. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—SAT Items (Forms D and E) to NAEP Framework
Assessment items = 108 (54 items per form)

	SAT Form D		**SAT Form E**	
	Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
Codable items	53	53	53	53
Uncodable items	1	1	1	1
Total assessment items	54	54	54	54

As seen in Table 9, both panels judged one SAT item to be uncodable to the NAEP objectives. The specific skills included in the uncodable item are discussed following Table 10.

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

Table 10. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—SAT Items (Forms D and E) to NAEP Framework
Assessment items = 108 (54 items per form)

	SAT Form D				**SAT Form E**			
	Panel 1		**Panel 2**		**Panel 1**		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**	**Mean Hits**	**Percentage**	**Mean Hits**	**Percentage**
Codable	53.38	97	53.50	98	52.88	98	53.38	98
Uncodable	1.88	3	1.00	2	1.13	2	1.13	2
Total	55.25		54.50		54.00		54.50	

For the 108 items (54 per form), the total mean hits for the two panels were 55.25 and 54.50 in Form D, and 54.00 and 54.50 in Form E. Where the numbers exceed 54, at least one item was coded to multiple objectives by one or more panelists. For Form D, panelists in Panel 1 assigned uncodable ratings to two different items and panelists in Panel 2 assigned an uncodable rating to one item. For Form E, panelists in Panel 1 assigned an uncodable rating to one item and panelists in Panel 2 assigned an uncodable rating to one item. One panelist in each panel assigned an uncodable rating to one additional item. Uncodable items included the skills of logical reasoning in real-world and algebraic contexts. No NAEP objectives are associated with these skills.

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

Table 11. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—SAT Items (Forms D and E) to NAEP Framework
Assessment items = 108 (54 items per form)

	SAT Form D						**SAT Form E**					
	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–Number properties and operations	12.13	23	22	13.00	24	24	11.38	22	21	11.63	22	21
2–Measurement	6.00	11	11	5.63	11	10	3.88	7	7	4.63	9	8
3–Geometry	10.75	20	19	11.00	21	20	9.25	17	17	10.00	19	18
4–Data analysis, statistics, and probability	5.50	10	10	5.50	10	10	6.63	13	12	6.25	12	11
5–Algebra	19.00	36	34	18.38	34	34	21.75	41	40	20.88	39	38
Total	53.38	100	97	53.50	100	98	52.88	100	98	53.38	100	98

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

All NAEP standards received hits from the SAT items. Of the five standards, Standard 5, “Algebra,” received the greatest number of mean hits across all panelists in Form D—19.00 mean hits in Panel 1 (36% of total hits) and 18.38 mean hits in Panel 2 (34% of total hits)—and in Form E—21.75 mean hits in Panel 1 (41% of total hits) and 20.88 mean hits in Panel 2 (39% of total hits). Standard 1, “Number properties and operations,” also received a substantial number of mean hits across both panels in Form D—12.13 mean hits in Panel 1 (23% of total hits) and 13.00 mean hits in Panel 2 (24% of total hits)—and in Form E—11.38 mean hits in Panel 1 (22% of total hits) and 11.63 mean hits in Panel 2 (22% of total hits). Standard 4, “Data analysis, statistics, and probability,” received the fewest mean hits in both panels in Form D—5.50 mean hits in Panel 1 (10% of total hits) and 5.50 mean hits in Panel 2 (10% of total hits)—and Standard 2, “Measurement,” received the fewest mean hits in both panels in Form E—3.88 mean hits in Panel 1 (7% of total hits) and 4.63 mean hits in Panel 2 (9% of total hits).

In comparison with the baseline alignment distribution of the NAEP short-version sample to the NAEP framework in Sub-Study 1, the distributions of SAT items to the NAEP framework were within 2 percentage points for Standard 1, “Number properties and operations”; Standard 4, “Data analysis, statistics, and probability”; and (for Form D) Standard 5, “Algebra.” For Form E, the emphasis on “Algebra” was within 7 percentage points of that in the baseline alignment. The

SAT items also had similar distribution for Standard 2, “Measurement,” and Standard 3, “Geometry,” combined (within 1 percentage point for Form D and 8 percentage points for Form E); however, while the ratio of “Measurement” items to “Geometry” items for the NAEP items was 1:1, the ratio for the SAT items was approximately 1:2.

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

Table 12. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—SAT Items (Forms D and E) to NAEP Framework
Assessment items = 108 (54 items per form)

			SAT Form D				**SAT Form E**			
			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	0.13	0	0	0	0	0	0	0
	1.1	1.1.d	1.13	2	0.5	1	1.63	3	0.38	1
		1.1.f	0	0	0.13	0	0	0	0	0
		1.1.g	1.25	2	0.63	1	0.63	1	1	2
		1.1.i	0	0	0.75	1	0.13	0	0.13	0
	1.2	1.2.b	0.13	0	0	0	0	0	0	0
		1.2.c	0	0	0.13	0	0	0	0	0
		1.2.d	0	0	0	0	0	0	0	0
		1.2	0.38	1	0	0	0	0	0	0
	1.3	1.3.a	0	0	0	0	0.13	0	0	0
		1.3.b	1	2	1.75	3	0.13	0	1.13	2
		1.3.c	0	0	0.25	0	0	0	0	0
		1.3.d	1	2	1	2	0.13	0	1.13	2
		1.3.f	0.88	2	2.63	5	2.13	4	2.13	4
		1.3	0.13	0	0	0	0	0	0	0
	1.4	1.4.c	1.38	3	2.13	4	3	6	2.5	5
		1.4.d	1.38	3	1.5	3	1	2	0.88	2
		1.4	0	0	0	0	0.25	0	0	0
	1.5	1.5.c	1	2	0.75	1	0.25	0	0.38	1
		1.5.d	0.75	1	0	0	0.75	1	1	2
		1.5.e	0.63	1	0.75	1	0	0	0.13	0
		1.5.f	0.25	0	0	0	0.13	0	0.88	2
		1.5	0	0	0	0	1	2	0	0
	1.6	1.6.a	0.75	1	0.13	0	0	0	0	0
		1.6.b	0	0	0	0	0	0	0	0
		1.6	0	0	0	0	0.13	0	0	0
2	2.1	2.1.b	0	0	0	0	0	0	0	0
		2.1.c	0	0	0	0	0	0	0	0
		2.1.d	2.5	5	3	6	2.13	4	2.25	4
		2.1.f	2.75	5	2.63	5	1.38	3	1.5	3
		2.1.h	0.5	1	0	0	0.13	0	0.38	1

			SAT Form D				SAT Form E			
			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
		2.1.i	0.13	0	0	0	0.13	0	0.38	1
		2.1	0	0	0	0	0.13	0	0	0
	2.2	2.2.a	0	0	0	0	0	0	0	0
		2.2.b	0	0	0	0	0	0	0	0
		2.2.d	0	0	0	0	0	0	0	0
		2.2.e	0	0	0	0	0	0	0	0
		2.2.f	0.13	0	0	0	0	0	0	0
	2.3	2.3.a	0	0	0	0	0	0	0	0
		2.3.b	0	0	0	0	0	0	0	0
		2.3.c	0	0	0	0	0	0	0	0
		2.3.d	0	0	0	0	0	0	0	0
		2.3.e	0	0	0	0	0	0	0	0
		2.3.f	0	0	0	0	0	0	0.13	0
		2.3.g	0	0	0	0	0	0	0	0
3	3.1	3.1.c	0	0	0	0	0	0	0	0
		3.1.d	0	0	0	0	0	0	0	0
		3.1.e	0.5	1	1.75	3	0	0	0	0
		3.1.f	0.13	0	0	0	0	0	0	0
	3.2	3.2.a	0	0	0	0	0.38	1	0.75	1
		3.2.b	0	0	0	0	0	0	0.13	0
		3.2.c	0	0	0	0	0	0	0	0
		3.2.d	0	0	0	0	0	0	0	0
		3.2.e	0	0	0	0	0	0	0	0
		3.2.g	0	0	0	0	0	0	0	0
		3.2	0	0	0	0	0.25	0	0	0
	3.3	3.3.b	3.63	7	2.38	4	2.75	5	2.25	4
		3.3.c	0	0	0	0	0.75	1	0	0
		3.3.d	1.63	3	1.5	3	0.88	2	2.13	4
		3.3.e	0.25	0	0.25	0	0.38	1	1.25	2
		3.3.f	0.25	0	0.75	1	0.13	0	0.13	0
		3.3.g	1.25	2	0.63	1	0.38	1	0.5	1
		3.3.h	0.75	1	1.13	2	1.38	3	0.38	1
	3.4	3.4.a	1.63	3	2.13	4	1.88	4	2.5	5
		3.4.b	0	0	0	0	0.13	0	0	0
		3.4.c	0	0	0	0	0	0	0	0
		3.4.d	0.63	1	0.5	1	0	0	0	0
		3.4.e	0	0	0	0	0	0	0	0
		3.4.f	0	0	0	0	0	0	0	0
		3.4.g	0	0	0	0	0	0	0	0
		3.4.h	0	0	0	0	0	0	0	0
		3.4	0.13	0	0	0	0	0	0	0
	3.5	3.5.a	0	0	0	0	0	0	0	0
		3.5.b	0	0	0	0	0	0	0	0

Standards	Goals	Objectives	SAT Form D				SAT Form E			
			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
		3.5.c	0	0	0	0	0	0	0	0
		3.5.d	0	0	0	0	0	0	0	0
		3.5.e	0	0	0	0	0	0	0	0
4		4	0	0	0	0	0.13	0	0	0
	4.1	4.1.a	2.63	5	2.25	4	1.88	4	1.5	3
		4.1.b	0.13	0	0.25	0	0	0	0.13	0
		4.1.c	0	0	0	0	0.13	0	0.13	0
		4.1.d	0	0	0	0	0	0	0	0
		4.1.e	0	0	0	0	0	0	0	0
		4.1.f	0	0	0	0	0	0	0	0
	4.2	4.2.a	1.13	2	1.13	2	1.63	3	2.25	4
		4.2.b	0	0	0	0	0.25	0	0.25	0
		4.2.c	0	0	0	0	1	2	0.13	0
		4.2.d	0	0	0	0	0	0	0.13	0
		4.2.e	0	0	0	0	0	0	0	0
		4.2.f	0	0	0	0	0	0	0	0
		4.2.g	0	0	0	0	0	0	0	0
	4.3	4.3.a	0	0	0	0	0	0	0	0
		4.3.b	0	0	0	0	0	0	0	0
		4.3.c	0	0	0	0	0	0	0	0
		4.3.d	0	0	0	0	0	0	0	0
		4.3.e	0	0	0	0	0	0	0	0
	4.4	4.4.a	0	0	0	0	0	0	0	0
		4.4.b	0.63	1	0.88	2	0.63	1	0.75	1
		4.4.c	0	0	0	0	0	0	0	0
		4.4.d	0	0	0	0	0	0	0	0
		4.4.e	0.88	2	0.88	2	0.75	1	0.88	2
		4.4.h	0	0	0	0	0.25	0	0.13	0
		4.4.i	0	0	0	0	0	0	0	0
		4.4.j	0.13	0	0	0	0	0	0	0
		4.4.k	0	0	0	0	0	0	0	0
	4.5	4.5.a	0	0	0	0	0	0	0	0
		4.5.b	0	0	0.13	0	0	0	0	0
		4.5.c	0	0	0	0	0	0	0	0
		4.5.d	0	0	0	0	0	0	0	0
		4.5.e	0	0	0	0	0	0	0	0
5	5.1	5.1.a	2.38	4	2.75	5	1.5	3	1.5	3
		5.1.b	0.25	0	0	0	0	0	0	0
		5.1.e	0.13	0	0.13	0	1.75	3	1.88	4
		5.1.g	0	0	0	0	0	0	0	0
		5.1.h	0	0	0	0	0	0	0.63	1
		5.1.i	0.13	0	0	0	0.13	0	0	0
		5.1.j	0	0	0	0	0	0	0	0

Standards	Goals	Objectives	SAT Form D				SAT Form E			
			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
	5.2	5.2.a	1.88	4	1.25	2	2.5	5	3.25	6
		5.2.b	0.13	0	0.13	0	0.38	1	0	0
		5.2.d	0	0	0	0	0.25	0	0	0
		5.2.e	0	0	0	0	0	0	0	0
		5.2.f	0	0	0	0	0	0	0.13	0
		5.2.g	0	0	0	0	0.13	0	0	0
		5.2.h	0	0	0	0	0	0	0	0
	5.3	5.3.b	1.5	3	1.13	2	2.38	4	2.25	4
		5.3.c	0.88	2	1.38	3	0.63	1	1.25	2
		5.3.d	0.25	0	1.63	3	3.13	6	2	4
		5.3.e	0.13	0	2.5	5	0.25	0	0	0
		5.3.f	1.63	3	2.13	4	1.88	4	1.25	2
		5.3.g	0	0	0.63	1	0	0	0	0
		5.3.h	0	0	0.13	0	0.13	0	0	0
	5.4	5.4.a	3.25	6	0.88	2	3.25	6	3.13	6
		5.4.c	1.13	2	0.25	0	1.88	4	1.63	3
		5.4.d	3.38	6	1.63	3	1.5	3	1.75	3
		5.4.e	0	0	0	0	0	0	0	0
		5.4.f	1.88	4	1.75	3	0.13	0	0	0
		5.4.g	0	0	0.13	0	0	0	0	0
	5.5	5.5.a	0.13	0	0	0	0	0	0	0
		5.5.b	0	0	0	0	0	0	0	0
		5.5.c	0	0	0	0	0	0	0	0
		5.5	0	0	0	0	0	0	0. 25	0

As shown in Table 12, 51 of the 130 objectives received one or more hits in at least one panel; additionally, items were coded directly to nine goals. Of the 31 objectives that received one or more hits in both panels on both forms, the following three objectives received the most hits in one panel from one form:

- 3.3.b, “Apply geometric properties and relationships to solve problems in two and three dimensions” (3.63 and 2.75 mean hits for Forms D and E in Panel 1 and 2.38 and 2.25 mean hits for Forms D and E in Panel 2)
- 5.4.a, “Solve linear, rational, or quadratic equations or inequalities, including those involving absolute value” (3.25 and 3.25 mean hits for Forms D and E in Panel 1 and 0.88 and 3.13 mean hits for Forms D and E in Panel 2)
- 5.4.d, “Solve (symbolically or graphically) a system of equations or inequalities and recognize the relationship between the analytical solution and graphical solution” (3.38 and 1.5 mean hits for Forms D and E in Panel 1 and 1.63 and 1.75 mean hits for Forms D and E in Panel 2)

The majority of panelists in at least one panel identified, across the two forms, seven items that were aligned to Objective 3.3.b, six items that were aligned to Objective 5.4.a, and six items that were aligned to Objective 5.4.d.

The items coded to Objective 3.3.b can be characterized as assessing student knowledge of geometric properties, with an emphasis on assessing the application of those concepts. The items coded to Objective 5.4.a and 5.4.d tend to be skills-based and involve solving equations and inequalities.

As shown in Table 12, the following 17 objectives had over two mean hits for at least one form in at least one panel:

- 1.3.f, "Solve application problems involving numbers, including rational and common irrationals"
- 1.4.c, "Use proportions to solve problems (including rates of change)"
- 2.1.d, "Solve problems of angle measure, including those involving triangles or other polygons or parallel lines cut by a transversal"
- 2.1.f, "Solve problems involving perimeter or area of plane figures such as polygons, circles, or composite figures"
- 3.3.b, "Apply geometric properties and relationships to solve problems in two and three dimensions"
- 3.3.d, "Use the Pythagorean theorem to solve problems in two- or three-dimensional situations"
- 3.4.a, "Solve problems involving the coordinate plane such as the distance between two points, the midpoint of a segment, or slopes of perpendicular or parallel lines"
- 4.1.a, "Read or interpret graphical or tabular representations of data"
- 4.2.a, "Calculate, interpret, or use summary statistics for distributions of data including measures of typical value (mean, median), position (quartiles, percentiles), and spread (range, interquartile range, variance, and standard deviation)"
- 5.1.a, "Recognize, describe, or extend numerical patterns, including arithmetic and geometric progressions"
- 5.2.a, "Create and translate between different representations of algebraic expressions, equations, and inequalities (e.g., linear, quadratic, exponential, or *trigonometric) using symbols, graphs, tables, diagrams, or written descriptions"
- 5.3.b, "Write algebraic expressions, equations, or inequalities to represent a situation"
- 5.3.d, "Write equivalent forms of algebraic expressions, equations, or inequalities to represent and explain mathematical relationships"
- 5.3.e, "Evaluate algebraic expressions including polynomials and rational expressions"
- 5.3.f, "Use function notation to evaluate a function at a specified point in its domain and combine functions by addition, subtraction, multiplication, division, and composition"
- 5.4.a, "Solve linear, rational, or quadratic equations or inequalities, including those involving absolute value"
- 5.4.d, "Solve (symbolically or graphically) a system of equations or inequalities and recognize the relationship between the analytical solution and graphical solution"

As can be seen from the preceding list, nine objectives across seven goals in Standards 1, 2, 3, and 4 had over two mean hits for at least one form in at least one panel; these objectives can be characterized as skills necessary to compute and solve problems involving real numbers, measures, geometric figures, and data. The eight objectives for Standard 5, “Algebra,” that had over two mean hits for at least one form in at least one panel require the student to demonstrate a basic knowledge of algebraic manipulations.

Within the five standards, ten goals received very few hits:

- 1.6, “Mathematical reasoning using number”
- 2.2, “Systems of measurement”
- 2.3, “Measurement in triangles”
- 3.1, “Dimension and shape”
- 3.2, “Transformation of shapes and preservation of properties”
- 3.4, “Position, direction, and coordinate geometry”
- 3.5, “Mathematical reasoning in geometry”
- 4.3, “Experiments and samples”
- 4.5, “Mathematical reasoning with data”
- 5.5, “Mathematical reasoning in algebra”

Important objectives included in these ten goals (e.g., objectives under Goals 1.6, 3.5, 4.5, and 5.5) involve certain higher-order thinking and reasoning skills such as those used to connect more than one objective and apply concepts to solve real-world and mathematical problems.

Of the 130 NAEP objectives, a total of 57 objectives received no hits by either panel for either form.

Comments recorded by the panelists expressed the observation that some important objectives were not addressed, that some of the codable items addressed only part of an objective, and that the items tended to address basic skills rather than assess conceptual development.

Tables 13a and 13b display 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 13a. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Eight Reviewers per Panel—SAT Items (Form D) to NAEP Framework
Assessment items = 54

	Alignment Criteria							
	Categorical Concurrence (mean hits)		**Depth-of-Knowledge Consistency** (% of hits at or above level of standard)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
1–Number properties and operations	12.12	13	77	87	48*	42*	0.86	0.79
2–Measurement	6	5.62**	65	62	15**	10**	0.86	0.9
3–Geometry	10.75	11	71	83	21**	22**	0.77	0.81
4–Data analysis, statistics, and probability	5.5**	5.5**	82	75	11**	12**	0.81	0.84
5–Algebra	19	18.38	76	92	31**	32**	0.79	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.

Table 13b. Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria as Rated by Eight Reviewers per Panel—SAT Items (Form E) to NAEP Framework
Assessment items = 54

	Alignment Criteria							
	Categorical Concurrence (mean hits)		**Depth-of-Knowledge Consistency** (% of hits at or above level of standard)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
1–Number properties and operations	11.38	11.62	96	75	34**	41*	0.77	0.8
2–Measurement	3.88**	4.62**	63	42*	12**	15**	0.87	0.82
3–Geometry	9.25	10	69	80	19**	19**	0.79	0.79
4–Data analysis, statistics, and probability	6.62	6.25	78	68	16**	14**	0.85	0.82
5–Algebra	21.75	20.88	72	62	34**	34**	0.78	0.82

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 108 SAT assessment items analyzed, all but one (107) were found to match or "hit" objectives. Using the typical WAT threshold value of 6 mean hits, for Form D, categorical concurrence was met for all standards except Standard 2, "Measurement," which had 5.62 mean hits in Panel 2, and Standard 4, "Data analysis, statistics, and probability," which had 5.5 mean hits in Panel 1. For Form E, categorical concurrence was met for all standards except Standard 2, “Measurement,” which had 3.88 and 4.62 mean hits in Panels 1 and 2, respectively. In comparing these results to the baseline alignment in Sub-Study 1, it is important to consider the difference in the number of items reviewed and compare the relative proportional distribution between the two tests.

Depth-of-knowledge consistency was met for “Number properties and operations,” “Geometry,” “Data analysis, statistics, and probability,” and “Algebra,” with more than 50% of the items at or above the DOK level of the standard to which they aligned. For Form E in Panel 2, depth-of-knowledge consistency was not met, as only 42% of items were at or above the DOK level of the objective to which they aligned; however, for Form D in Panel 2, and for both forms in Panel 1, this criterion was met. Most of the items in the set of 108 items were coded to DOK Level 2, although items were coded to DOK Levels 1 and 3 as well. Discrepancies of greater than five percentage points exist between the panels for this criterion for all five standards, with the exception of “Measurement” for Form D. Although WAT system server errors prevented a comparison of overall reports for this criterion during the workshop, facilitators compared the item-level codes during cross-panel adjudication, and differences in item-level DOK codes were identified. The facilitators identified items with the greatest differences in DOK coding by the two panels and discussed them with their panels. These items tended to be coded at Level 1 or Level 2, with a distribution of codes across panelists in each panel; the difference was not found to be systematic. Following the discussion, panelists were given the opportunity to change codes.

Range of knowledge was not met for any of the five standards of the NAEP framework in either Form D or Form E. The limited range of knowledge found in the SAT items is consistent with differences in the frameworks of the two tests: the SAT framework comprises a list of a broad collection of skills, while the NAEP framework includes objectives that involve higher-level thinking and the application of concepts. Therefore, one would expect SAT to cover a narrower range of concepts than the NAEP objectives require.

The balance of representation criterion (a WAT threshold of 0.7) was met for all five standards in both Form D and Form E. This indicates that the items were relatively evenly distributed among those objectives that received hits.

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

In Sub-Study 3, reviewers evaluated the alignment between the SAT items and the SAT framework. A short-form sample of 40 items from the SAT mathematics test Form D was analyzed. The results of Sub-Study 3 are presented in Tables 14–19.

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

Table 14. Codability of Items as Determined by Items Rated Uncodable by Eight Reviewers per Panel—SAT Items (Short Version) to SAT 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 14, all 40 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 15 displays the numbers and percentages of mean hits assigned to items by each panel. Codable mean hits are the total hits to objectives, divided by the number of reviewers. Uncodable mean hits are the number of uncodable ratings assigned, divided by the number of reviewers.

Table 15. Number and Percentage of Mean Hits (Codable and Uncodable) as Rated by Eight Reviewers per Panel—SAT Items (Short Version) to SAT Framework
Assessment items = 40

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	41.63	100%	40.25	100%
Uncodable	0.00	0%	0.00	0%
Total	41.63		40.25	

For the 40 SAT items, the total mean hits for the two panels were 41.63 and 40.25. The numbers exceed 40 because some items were coded to multiple objectives by at one or more panelists. No uncodable ratings were assigned.

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

Table 16. Categorical Concurrence between Standards and Assessment as Rated by Eight Reviewers per Panel—SAT Items (Short Version) to SAT 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**
N–Number and operations	9.25	22	22	8.00	20	20
A–Algebra and functions	14.50	35	35	15.00	37	37
G–Geometry and measurement	14.13	34	34	13.63	34	34
D–Data analysis, statistics, and probability	3.75	9	9	3.63	9	9
Total	41.63	100	100	40.25	100	100

All SAT standards received hits from SAT items in the short-version subset, with a distribution as shown in Table 16. Of the four standards, SAT A, “Algebra and functions,” received the greatest number of mean hits in both panels (14.50 and 15.00, respectively), making up 35% and 37% of total hits, respectively. SAT G, “Geometry and measurement,” received the second greatest number of mean hits in both panels: 14.13 mean hits in Panel 1 (34% of total hits) and 13.63 mean hits in Panel 2 (34% of total hits). SAT D, “Data analysis, statistics, and probability,” received the fewest mean hits in both panels: 3.75 mean hits in Panel 1 (9% of total hits) and 3.63 mean hits in Panel 2 (9% of total hits).

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

Table 17. Number and Percentage of Mean Hits to Objectives as Rated by Eight Reviewers per Panel—SAT Items (Short Version) to SAT Framework
Assessment items = 40

Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
N–Number and operations	N.1	1.88	5	1.25	3
	N.2	0.25	1	0	0
	N.3	1.25	3	1.38	3
	N.4	2.25	5	2.25	6
	N.5	0.5	1	0	0
	N.6	0.75	2	1	2
	N.7	2.38	6	2.13	5
A–Algebra and functions	A.1	1.63	4	2.25	6
	A.2	4.75	11	3.75	9
	A.3	1.5	4	1.25	3
	A.4	1.63	4	2	5

Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
	A.5	1	2	0.88	2
	A.6	0	0	0	0
	A.7	0.75	2	1	2
	A.8	0.88	2	0.75	2
	A.9	0.13	0	0.88	2
	A.10	2.25	5	2.25	6
G–Geometry and measurement	G.1	1.38	3	1	2
	G.2	1.38	3	1.13	3
	G.3	1.88	5	2	5
	G.4	2.38	6	1.63	4
	G.5	2.63	6	3	7
	G.6	1.25	3	1	2
	G.7	0.13	0	0.13	0
	G.8	0.75	2	1.25	3
	G.9	2.38	6	2.5	6
D–Data analysis, statistics, and probability	D.1	1.63	4	1.5	4
	D.2	0.88	2	1.13	3
	D.3	1.25	3	1	2

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

As shown in Table 17, 28 of the 29 objectives received one or more hits in both panels from the SAT short-version items. Objective A.2, “Algebraic representations, translation, and algebraic word problems (those that usually require an algebraic equation to solve),” had the greatest number of mean hits in both panels (4.75 mean hits in Panel 1 and 3.75 mean hits in Panel 2). A majority of panelists across both panels identified two items that were aligned to Objective A.2; these items can be characterized as involving algebraic manipulations in a problem-solving context (real-world or mathematical).

As shown in Table 17, the following eight objectives had over two mean hits in at least one panel:

- N.4, “Sequences and series”
- N.7, “Logic/logical reasoning”
- A.1, “Operations with real numbers”
- A.2, “Algebraic representations, translation, and algebraic word problems”
- A.10, “Basic concepts of algebraic functions”
- G.4, “Special triangles (30-60-90, isosceles, equilateral, etc.)”
- G.5, “Circles”
- G.9, “Coordinate geometry”

Two objectives received few hits in comparison with all the other objectives: Objective N.2, “Rational numbers,” received 0.25 mean hits from Panel 1 and 0 mean hits from Panel 2, and Objective G.7, “Solid geometric figures,” received 0.13 mean hits from both Panel 1 and Panel 2. Finally, Objective A.6, “Radical equations,” received no hits from either panel.

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

Table 18. Summary of Attainment of Acceptable Alignment Level on Three Content Focus Criteria as Rated by Eight Reviewers per Panel—SAT Items (Short Version) to SAT Framework
Assessment items = 40

	Alignment Criteria					
	Categorical Concurrence (mean hits)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
N–Number and operations	9.25	8	71	70	0.81	0.82
A–Algebra and functions	14.5	15	65	79	0.76	0.76
G–Geometry and measurement	14.12	13.62	81	86	0.84	0.8
D–Data analysis, statistics, and probability	3.75**	3.62**	92	92	0.9	0.88

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 SAT items analyzed, all (40) were found to match objectives, or have “hits.” Using the typical WAT threshold value of 6 mean hits, categorical concurrence was met for all SAT standards except for SAT D, “Data analysis, statistics, and probability.”

Range of knowledge was met by all four standards, each of which had over 50% of its objectives hit.

Balance of representation was met by all four standards as well, indicating that for the objectives receiving hits, the distribution of items across those objectives was well distributed.

As described earlier, the SAT framework was not able to be coded for DOK. Table 19 shows the range of DOK of the SAT items aligned to each SAT standard.

Table 19. Range of Depth of Knowledge of SAT Items (Short Version) Aligned to the SAT Framework
Assessment Items = 40

	Panel 1							Panel 2						
		DOK Level 1		DOK Level 2		DOK Level 3			DOK Level 1		DOK Level 2		DOK Level 3	
Standards	**Mean Hits**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**
N–Number and operations	9.25	2.25	23	5.88	66	1.13	10	8.00	0.88	9	6.25	82	0.88	9
A–Algebra and functions	14.50	4.38	30	9.50	65	0.63	4	15.00	2.00	13	12.50	83	0.50	3
G–Geometry and measurement	14.13	4.75	34	9.00	64	0.38	3	13.63	2.63	19	10.38	76	0.63	5
D–Data analysis, statistics, and probability	3.75	1.00	28	2.25	59	0.50	14	3.63	1.00	28	2.25	61	0.38	11
Total	41.63	12.38	30	26.63	64	2.63	5	40.25	6.50	17	31.38	77	2.38	5

As shown in Table 19, the majority of the items were coded at DOK Level 2 by both panels, although to different degrees in terms of percentage of mean hits to the standard (64% by Panel 1 and 77% by Panel 2). Based on percentage of mean hits, Panel 1 found 30% to be at Level 1, while Panel 2 found 17% to be at Level 1. A smaller percentage of items were found to be at DOK Level 3 (5% of mean hits by both panels).

Similarly, in each of the four standards, the majority of mean hits were from DOK Level 2 items. For items aligned to “Number and operations,” 66% of mean hits in Panel 1 and 82% of mean hits in Panel 2 were for DOK Level 2 items. For items aligned to “Algebra,” the percentages of mean hits from Level 2 items were 65% and 83%. For items aligned to “Geometry and measurement,” the percentages of mean hits from Level 2 items were 59% and 61%. For items aligned to “Data analysis, statistics, and probability,” the percentages of mean hits from Level 2 items were 64% and 76%.

The differences of greater than five percentage points exist in the results for DOK Levels 1 and 2 for three of the four standards. A review of the item-level coding during cross-panel adjudication indicated a number of items for which the panelists did not have agreement in the DOK ratings. Generally, Panel 2 coded a greater percentage of items at Level 2 than did Panel 1. In practice, DOK levels are not absolute, and some of the items in question could reasonably be interpreted at either “high” Level 1 or “low” Level 2. The interpretation of the items was discussed with panelists and they were given an opportunity to change ratings. Given the relatively small numbers of items and mean hits causing the percentage differences at each standard, the discrepancy was determined not to be systematic. Overall, the panels were found to be replicate, as defined by the study.

Sub-Study 4—NAEP Items to SAT Framework

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

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

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

	Panel 1	**Panel 2**
Codable items	161	162
Uncodable items	3	2
Total assessment items	164	164

As shown in Table 20, Panel 1 identified three NAEP items as uncodable to the SAT framework. Panel 2 identified two items as uncodable to the SAT framework. The specific skills included in the uncodable items are discussed following Table 21.

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

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

	Panel 1		**Panel 2**	
	Mean Hits	**Percentage**	**Mean Hits**	**Percentage**
Codable	160.38	96%	163.50	98%
Uncodable	7.25	4%	3.38	2%
Total	167.63		166.88	

For the 164 items, the total mean hits for each panel were 167.63 and 166.88, and the codable mean hits for the two panels were 160.38 and 163.50. There were a number of uncodable ratings in each panel, comprising 4% and 2% of mean hits in Panel 1 and Panel 2, respectively. In addition to the items rated uncodable by all panelists, three items were found to be uncodable by the majority (5 or more of 8) of panelists in Panel 1, and additional items were rated uncodable by one or more panelists in either panel. Of the five items rated uncodable by the majority of panelists in either panel, three items involve trigonometry. The other two items were part of a cluster that has a spreadsheet as a stimulus.

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

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

	Panel 1			Panel 2		
Standards	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**	**Mean Hits**	**% of Total Hits**	**% of Hits Adjusted for Uncodable**
N–Number and operations	31.13	19	19	34.13	21	20
A–Algebra and functions	52.75	33	31	50.50	31	30
G–Geometry and measurement	42.63	27	25	44.75	27	27
D–Data analysis, statistics, and probability	33.88	21	20	34.13	21	20
Total	160.38	100	96	163.50	100	98

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

All SAT standards received hits from NAEP items, with the distribution shown in Table 22. Of the four objectives, SAT A, “Algebra and functions,” received the greatest number of mean hits in both panels (52.75 and 50.50, respectively), making up 33% and 31% of the item set, respectively. SAT N, “Number and operations,” received the fewest mean hits in Panel 1—31.13 mean hits (19% of total hits)—and both “Number and operations” and SAT D, “Data analysis, statistics, and probability,” received the fewest mean hits in Panel 2: 34.13 mean hits (21% of total hits). These results are similar to the SAT items–to–SAT objectives results in Sub-Study 3, in that the greatest number of items in both tests were found to align to “Algebra and functions;” however, the NAEP items had proportionally less emphasis on “Geometry and measurement” and more emphasis on “Data analysis, statistics, and probability” than did the SAT items.

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

		Panel 1		**Panel 2**	
Standards	**Objectives**	**Mean Hits**	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
N–Number and operations	N	1.5	1	0.13	0
	N.1	3.63	2	3	2
	N.2	4.38	3	3.75	2
	N.3	12.75	8	15.88	10
	N.4	3.38	2	4.5	3
	N.5	0.25	0	0.88	1
	N.6	2.13	1	1.88	1
	N.7	3.13	2	4.13	3
A–Algebra and functions	A	6.63	4	7.88	5
	A.1	6	4	7.88	5
	A.2	9.63	6	9.13	6
	A.3	2.88	2	1.25	1
	A.4	1.25	1	2.25	1
	A.5	1.13	1	1.13	1
	A.6	0.13	0	0.13	0
	A.7	2.13	1	2.38	1
	A.8	3.75	2	2.75	2
	A.9	0.13	0	0.13	0
	A.10	19.13	12	15.63	10
G–Geometry and measurement	G	7.75	5	5.88	4
	G.1	2.5	2	3.5	2
	G.2	1.75	1	1.25	1
	G.3	0.88	1	0.88	1
	G.4	6.75	4	8.88	5
	G.5	3.88	2	3.63	2
	G.6	4.75	3	2.25	1
	G.7	2.5	2	4.5	3
	G.8	6.5	4	5.88	4
	G.9	5.38	3	8.13	5
D–Data analysis, statistics, and probability	D	8.13	5	2	1
	D.1	13.5	8	14	9
	D.2	3.5	2	8.13	5
	D.3	8.75	5	10	6

As shown in Table 23, all 29 objectives received one or more hits in both panels. Three objectives had the greatest number of mean hits in both panels:

- N.3, “Arithmetic algebraic word problems”
- A.10, “Basic concepts of algebraic functions”
- D.1, “Data interpretation”

A majority of panelists across both panels identified 14 items that were aligned to Objective N.3, 16 items that were aligned to Objective A.10, and 15 items that were aligned to Objective D.1.

As shown in Table 23, the following seven objectives had over six mean hits in both panels:

- N.3, “Arithmetic algebraic word problems”
- A.1, “Operations with real numbers”
- A.2, “Algebraic representations, translation, and algebraic word problems”
- A.10, “Basic concepts of algebraic functions”
- G.4, “Special triangles”
- D.1, “Data interpretation”
- D.3, “Probability”

In addition, three of the four standards (“Algebra and functions, “Geometry and measurement,” and “Data analysis, statistics, and probability”) received between 1% and 5% of total hits at the standard level rather than at the objective level.

Four objectives received fewer than 1% of the total mean hits across both panels:

- N.5, “Sets”
- A.6, “Radical equations”
- A.9, “Direct and inverse variation”
- G.3, “Triangles (nonspecial)”

The comments recorded by the panelists observed that the SAT objectives were in fact broad topic statements and that some found it difficult to code the NAEP items to them.

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

Table 24. Summary of Attainment of Acceptable Alignment Level on Three Content Focus Criteria as Rated by Eight Reviewers per Panel—NAEP Items to SAT Framework
Assessment items = 164[15]

	Categorical Concurrence (mean hits)		**Range of Knowledge** (% of objectives hit)		**Balance of Representation** (balance index)	
Standards	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
N–Number and operations	31.12	34.12	91	95	0.7	0.67*
A–Algebra and functions	52.75	50.5	82	80	0.64*	0.65*
G–Geometry and measurement	42.62	44.75	95	96	0.74	0.75
D–Data analysis, statistics, and probability	33.88	34.12	100	100	0.79	0.78

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

Of the 164 NAEP assessment items analyzed, not all were found to match or “hit” objectives. As previously noted, there were a number of items found by panelists to be uncodable. However, using the typical WAT threshold value of six mean hits, categorical concurrence was met for all four of the SAT standards, with between 31.12 and 52.75 mean hits to the standards.

The range of knowledge criterion was met above the 50% threshold for all four of the SAT standards, indicating that over 50% of the objectives in each standard received hits.

The balance of representation criterion, a WAT threshold of 0.7 was weakly met for SAT A, “Algebra and functions,” with 0.64 and 0.65 mean hits for Panel 1 and Panel 2, respectively. Balance of representation was met for all other standards.

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

[15] The percentages in this table indicate the distribution of total hits. It should be noted that, as shown in Table 21, 4% and 2% of the adjusted total hits for NAEP items were determined by panelists to be uncodable to any objective.

Table 25. Range of Depth of Knowledge of NAEP Items Aligned to the SAT Framework
Assessment Items = 164

	Panel 1							**Panel 2**						
		DOK Level 1		**DOK Level 2**		**DOK Level 3**			**DOK Level 1**		**DOK Level 2**		**DOK Level 3**	
Standards	**Mean Hits**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**	**Mean Hits**	**% of Mean Hits to Std.**
N–Number and operations	31.13	13.25	43	15.25	49	2.50	9	34.13	8.13	24	22.38	65	3.63	11
A–Algebra and functions	52.75	27.88	53	23.75	45	1.13	2	50.50	19.13	38	30.13	60	1.25	2
G–Geometry and measurement	42.63	14.38	34	24.25	57	3.88	9	44.75	10.13	23	29.25	66	5.38	12
D–Data analysis, statistics, and probability	33.88	13.38	40	19.13	56	1.38	4	34.13	4.63	14	26.88	78	2.63	8
Total	160.38	68.88	46	82.38	47	8.88	7	163.50	42.00	27	108.63	61	12.88	12

As shown in Table 25, most of the items were found to be at DOK Level 2 by both panels, although to different degrees in terms of percentage of mean hits to the standard (47% in Panel 1 and 61% in Panel 2). Based on the percentage of mean hits, both panels found a substantial number of the items to be at DOK Level 1 (46% in Panel 1 and 27% in Panel 2). A smaller percentage of items were found to be at DOK Level 3 (7% in Panel 1 and 12% in Panel 2).

Similarly, in each of the four standards except SAT A, “Algebra and functions,” in Panel 1, the highest percentage of mean hits were from DOK Level 2 items. For items aligned to SAT N, “Number and operations,” 49% of mean hits in Panel 1 and 65% of mean hits in Panel 2 were for DOK Level 2 items. For items aligned to “Algebra and functions,” the percentages of mean hits from Level 2 items were 45% and 60%, and from Level 1 items, 53% and 38%. For items aligned to “Geometry and measurement,” the percentages of mean hits from Level 2 items were 57% and 66%. For items aligned to “Data analysis, statistics, and probability,” the percentages of mean hits from Level 2 items were 56% and 78%.

Differences between panels of greater than five percentage points in the results for DOK Levels 1 and 2 are apparent for all four standards. As in the previous sub-study, a review of the item-level coding during cross-panel adjudication indicated a number of items for which the panelists did not have agreement on the DOK ratings. Generally, Panel 2 coded a greater percentage of items at Level 2 than did Panel 1. However, the differences were not found to be the result of any systematic difference in application of the criteria between the panels, but instead appeared to be a combination of smaller differences in the distribution in each panel of panelists’ judgments across adjacent DOK levels on individual items. The interpretation of the items was discussed with panelists and, although agreement across panels was not required by the study design for results to be valid, panelists were given an opportunity to change ratings if their judgments of the items had changed.

In comparison to the baseline alignment of the SAT short-version to the SAT framework, NAEP had a slightly wider range of depth of knowledge. For all standards, the majority of SAT items were coded to DOK Level 2, whereas NAEP items for all standards but “Geometry” and “Measurement” were more evenly distributed between DOK Level 1 and 2. Both tests had a low percentage of items coded to DOK Level 3 for all standards.

IV. Panelists' Evaluations of the Process

This section details the findings from responses to training and process evaluation questionnaires that the eight panelists from each of two panels (a total of 16 panelists) completed at the end of each day of participation. WestEd 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 26 shows results for selected-response questions 5–9, 12, and 13, by panel. Numbers in bold font represent the highest number of responses for each question, by panel.

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

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

	Panel 1 (n=8)			**Panel 2** (n=8)		
How well did your facilitator…	**Not Well**	**Moderately Well**	**Very Well**	**Not Well**	**Moderately Well**	**Very Well**
Q13. facilitate today's consensus process?	0	2	**6**	0	1	**7**

*n=7 for this question.

As shown in Table 26, all but one panelist across the two panels reported that the introductory session either adequately or very well explained the purposes of the study and introduced the NAEP and SAT assessments; one panelist in Panel 2 commented that the alignment questions were too vague. All panelists across the two panels reported that the introductory session either adequately or very well prepared them for understanding definitions of the DOK levels and using the WAT system. All but one panelist reported that the day’s training adequately or very well prepared them for the discussion process that led to agreement on DOK levels for NAEP objectives across the two panels; one panelist from Panel 1 felt only somewhat prepared and mentioned that while the group ultimately came to an understanding of the consensus process, the process was difficult at times.

When asked how comfortable they felt with the process of assigning DOK levels to objectives, the majority of panelists on both panels reported feeling either comfortable or very comfortable. One panelist from Panel 1 reported feeling somewhat comfortable with the process, but did not elaborate. All panelists across both panels felt the facilitators managed the discussions that led to agreement on DOK levels for NAEP objectives across the two panels moderately well or 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 suggested that providing more information regarding the verbs (e.g., “analyze,” “model”) used in the materials would be helpful, as would establishing agreement on the DOK levels of such verbs. Recommendations for improving the training and alignment process included providing more examples for each DOK level to better understand the nuances of coding decisions and clarifying the rubric for 12th grade. When asked what aspects of the day went particularly well, four of the 16 panelists mentioned the discussion process; two panelists mentioned the facilitator, while two panelists responded that everything went well. Five panelists from Panel 1 voiced concern over the length of the working day.

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. 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 27 shows results for selected-response questions 4, 5, and 6, by panel. Numbers in bold font represent the highest number of responses for each question, by panel.

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

<table>
<tr><th></th><th colspan="6"><b>Panel 1</b> (n=8)</th><th colspan="6"><b>Panel 2</b> (n=8)</th></tr>
<tr><th><b>How well did the training…</b></th><th><b>Not Well</b></th><th colspan="2"><b>Some-what</b></th><th colspan="2"><b>Ade-quately</b></th><th><b>Very Well</b></th><th><b>Not Well</b></th><th colspan="2"><b>Some-what</b></th><th colspan="2"><b>Ade-quately</b></th><th><b>Very Well</b></th></tr>
<tr><td>Q4. prepare you to assign DOK levels to test items?</td><td>0</td><td colspan="2">0</td><td colspan="2">2</td><td><b>6</b></td><td>0</td><td colspan="2">0</td><td colspan="2"><b>5</b></td><td>3</td></tr>
<tr><td>Q5. prepare you for the alignment (coding) process?</td><td>0</td><td colspan="2">0</td><td colspan="2">3</td><td><b>5</b></td><td>0</td><td colspan="2">0</td><td colspan="2"><b>6</b></td><td>2</td></tr>
<tr><td><b>How well did your facilitator…</b></td><td colspan="2"><b>Not Well</b></td><td colspan="2"><b>Moderately Well</b></td><td colspan="2"><b>Very Well</b></td><td colspan="2"><b>Not Well</b></td><td colspan="2"><b>Moderately Well</b></td><td colspan="2"><b>Very Well</b></td></tr>
<tr><td>Q6. facilitate today's consensus process?</td><td colspan="2">0</td><td colspan="2">0</td><td colspan="2"><b>8</b></td><td colspan="2">0</td><td colspan="2">1</td><td colspan="2"><b>7</b></td></tr>
</table>

All panelists reported feeling either adequately or very well prepared to both assign DOK levels to items and code items to objectives. All panelists also reported that the facilitator conducted the within-panel discussions either moderately well or very well.[16]

Recommendations for improving the alignment process or requests for more information included providing more examples during the training, more practice with the SAT framework before coding items to the framework, and more group discussion time (including discussion of decision rules) prior to independent coding. Two panelists (13%) also mentioned that there was too much work to be completed within the day. When asked what activities went particularly well, four panelists indicated the within-panel discussions, two panelists indicated the entire process, one panelist indicated the item DOK coding, and one panelist indicated the facilitator. No panelists indicated any areas in which they felt unprepared or wanted more information.

Day 3 Process Evaluation

The third day of the study comprised the review of the SAT framework objectives and mapping of SAT 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 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 seven responses are reported in each panel.

[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.

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

<table>
<tr><th></th><th colspan="12">**Panel 1** (n=7)</th><th colspan="12">**Panel 2** (n=7)</th></tr>
<tr><th>**How comfortable do you feel…**</th><th colspan="3">**Uncom-fortable**</th><th colspan="3">**Some-what**</th><th colspan="3">**Com-fortable**</th><th colspan="3">**Very Com-fortable**</th><th colspan="3">**Uncom-fortable**</th><th colspan="3">**Some-what**</th><th colspan="3">**Com-fortable**</th><th colspan="3">**Very Com-fortable**</th></tr>
<tr><td>Q4. assigning DOK levels to test items?</td><td colspan="3">0</td><td colspan="3">0</td><td colspan="3">3</td><td colspan="3">**4**</td><td colspan="3">0</td><td colspan="3">0</td><td colspan="3">**4**</td><td colspan="3">3</td></tr>
<tr><td>Q5. aligning test items to objectives?</td><td colspan="3">0</td><td colspan="3">1</td><td colspan="3">**5**</td><td colspan="3">1</td><td colspan="3">0</td><td colspan="3">0</td><td colspan="3">**6**</td><td colspan="3">1</td></tr>
<tr><td>**How well did your facilitator…**</td><td colspan="4">**Not Well**</td><td colspan="4">**Moderately Well**</td><td colspan="4">**Very Well**</td><td colspan="4">**Not Well**</td><td colspan="4">**Moderately Well**</td><td colspan="4">**Very Well**</td></tr>
<tr><td>Q6. facilitate today’s consensus process?</td><td colspan="4">0</td><td colspan="4">0</td><td colspan="4">**7**</td><td colspan="4">0</td><td colspan="4">1</td><td colspan="4">**6**</td></tr>
</table>

On Day 3 of the workshop, all panelists across both panels continued to report feeling comfortable or very comfortable assigning DOK levels to test items, and all but one panelist in Panel 1 reported feeling comfortable or very comfortable aligning items to objectives. The panelist who responded that he/she felt somewhat comfortable provided no further explanation. All panelists indicated that the facilitators either moderately well or very well managed the day’s within-panel discussions.

Panelists were asked 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 panelists reported technical issues with the WAT system that impeded the alignment activities, and four panelists indicated that they would have liked more time to complete the day’s alignment activities. Two other panelists reported that they would have preferred to code all items within a section prior to the within-panel discussions, while another suggested aligning all items to one framework before moving on to the second framework in order to avoid confusion between the NAEP and SAT objectives. Overall, however, panelists again expressed positive feelings about the within-group discussions.

Day 4 Process Evaluation

The fourth day of the study comprised mapping of NAEP items to the SAT framework and beginning coding SAT items to the NAEP framework. At the end of the day, panelists were administered a process evaluation questionnaire that solicited feedback on these alignment activities. Table 29 shows results for selected-response questions 4, 5, and 6, by panel. Numbers in bold font represent the highest number of responses for each question, by panel. Two panelists in each panel did not complete this questionnaire; therefore, only six responses are reported in each panel.

Table 29. Panelist Responses to Day 4 Process Evaluation Questionnaire

<table>
<tr><th></th><th colspan="12">**Panel 1** (n=6)</th><th colspan="12">**Panel 2** (n=6)</th></tr>
<tr><th>How comfortable do you feel…</th><th colspan="3">Uncom-fortable</th><th colspan="3">Some-what</th><th colspan="3">Com-fortable</th><th colspan="3">Very Com-fortable</th><th colspan="3">Uncom-fortable</th><th colspan="3">Some-what</th><th colspan="3">Com-fortable</th><th colspan="3">Very Com-fortable</th></tr>
<tr><td>Q4. assigning DOK levels to test items?</td><td colspan="3">0</td><td colspan="3">1</td><td colspan="3">1</td><td colspan="3">**4**</td><td colspan="3">0</td><td colspan="3">0</td><td colspan="3">**4**</td><td colspan="3">2</td></tr>
<tr><td>Q5. aligning test items to objectives?</td><td colspan="3">0</td><td colspan="3">2</td><td colspan="3">1</td><td colspan="3">**3**</td><td colspan="3">0</td><td colspan="3">0</td><td colspan="3">**3**</td><td colspan="3">**3**</td></tr>
<tr><th>How well did your facilitator…</th><th colspan="4">Not Well</th><th colspan="4">Moderately Well</th><th colspan="4">Very Well</th><th colspan="4">Not Well</th><th colspan="4">Moderately Well</th><th colspan="4">Very Well</th></tr>
<tr><td>Q6. facilitate today's consensus process?</td><td colspan="4">0</td><td colspan="4">0</td><td colspan="4">**6**</td><td colspan="4">0</td><td colspan="4">1</td><td colspan="4">**5**</td></tr>
</table>

On Day 4, the majority of panelists continued to feel comfortable or very comfortable assigning DOK levels to test items. One panelist on Panel 1 reported feeling only somewhat comfortable doing so; this panelist also reported feeling only somewhat comfortable aligning test items to objectives, commenting that having additional time to revisit the NAEP framework would have been helpful. The second panelist from Panel 1 who reported feeling somewhat comfortable aligning test items to objectives did not elaborate as to why. One panelist in Panel 2 indicated that the facilitator managed the within-panel discussions moderately well, without leaving any additional comment; all other panelists reported that the facilitators managed the 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. As on Day 3, several panelists indicated a need for more time in order to complete the alignment activities without feeling rushed, although one panelist reported that the midday break was very helpful in preparing for the afternoon activities. Another panelist suggested that uncodable items be identified in advance of the alignment sessions, so that panelists do not have to individually struggle with them. As on the previous three days, panelists did not report any activities for which they felt particularly unprepared.

Day 5 End-of-Study Evaluation

On the final day of the study, panelists completed the mapping of SAT items to the NAEP framework and responded to additional questions about the alignment process, the effectiveness of their panels and facilitators, and the 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 30 shows results for selected-response questions 4–11 and 15, by panel. Numbers in bold font represent the highest number of responses for each question, by panel.

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

	Panel 1 (n=8)				**Panel 2** (n=8)			
How well did Monday's training prepare you…	**Not Well**	**Somewhat**	**Adequately**	**Very Well**	**Not Well**	**Somewhat**	**Adequately**	**Very Well**
Q4. for understanding DOK levels?	0	0	**4**	**4**	0	0	**6**	2
Q7. for the consensus process?	0	0	3*	**4***	0	0	**5**	3
Q8. for the alignment (coding) process?	0	0	3	**5**	0	0	**6**	2
How comfortable did you feel…	**Uncomfortable**	**Somewhat**	**Comfortable**	**Very Comfortable**	**Uncomfortable**	**Somewhat**	**Comfortable**	**Very Comfortable**
Q5. assigning DOK levels to objectives?	0	1	3	**4**	0	0	**8**	0
Q6. assigning DOK levels to test items?	0	0	3	**5**	0	0	**5**	3
How useful was/were…	**Not Useful**	**Somewhat Useful**	**Adequately Useful**	**Very Useful**	**Not Useful**	**Somewhat Useful**	**Adequately Useful**	**Very Useful**
Q9. information provided prior to the study?	0	1	**4**	3	1	2	**4**	1
Q10. on-site training and coding materials?	0	0	1	**7**	0	1	2	**5**
How qualified was your panel…	**Not Qualified**	**Somewhat Qualified**	**Adequately Qualified**	**Very Qualified**	**Not Qualified**	**Somewhat Qualified**	**Adequately Qualified**	**Very Qualified**
Q11. to conduct this type of alignment?	0	0	1	**7**	0	0	0	**8**
How easy was it…	**Not Easy**	**Somewhat Easy**	**Adequately Easy**	**Very Easy**	**Not Easy**	**Somewhat Easy**	**Adequately Easy**	**Very Easy**
Q15. to use the WAT for the alignment process?	0	0	2*	**5***	1	0	3	**4**

*n=7 for this question.

All panelists across the two panels reported feeling either adequately or very well prepared by the Day 1 training for DOK coding, for the within-panel discussions, and for the alignment process. Most panelists (15) were comfortable or very comfortable assigning DOK levels to objectives, with only one panelist reporting feeling only somewhat prepared; this panelist provided no further comment. All panelists reported feeling comfortable or very comfortable aligning DOK levels to test items. Twelve of the panelists felt that information provided to them prior to the study was adequately useful or useful, while four panelists across the two panels felt the information was not useful or only somewhat useful; these four panelists did not provide further comment. The majority of panelists (15) felt that the on-site training and coding materials were adequately or very useful; one panelist felt the on-site training and coding materials were somewhat useful, but did not elaborate.

All but one panelist reported that the panels were very qualified to conduct the alignment activities; the remaining panelist reported that the panel was adequately qualified. One panelist suggested including more teachers who teach math at the grade 12 level.

Fourteen panelists across both panels reported that the WAT system was adequately or very easy to use during the alignment process. One panelist reported that the system was not easy to use and commented on a specific aspect of the WAT's functionality. Other panelists commented on minor technical problems, such as timing out.

When asked to provide qualitative feedback about their facilitator's effectiveness in managing the within-panel discussions, all 16 panelists reported that the facilitators were very effective, providing guidance and engaging the entire group.

More substantive issues were addressed in an open-ended format. Generally, panelists reported that the alignment criteria were useful in capturing important aspects of each assessment item/objective, although some found the alignment process to be challenging. Two panelists mentioned that it was particularly difficult to determine alignment of items with greater complexity. Another commented that the criteria were too vague, which made the process somewhat subjective.

Of the 13 panelists who responded to the question, 11 reported that the alignment process effectively captured content *similarities* between the assessments. Comments from the remaining two panelists were that the process captured the differences better than the similarities, and that the process was too vague to capture the similarities. Only one panelist reported on the actual similarities between the two assessments, commenting that they shared similar item types and content overlap. The remaining panelists focused on the *differences* between the assessments. More specifically, of the 13 panelists who responded, 12 reported that the alignment process effectively captured content differences between the assessments. One panelist commented that the assessments differed in terms of depth and number of topics, the complexity of the mathematics, and the response requirements. Other panelists mentioned that NAEP was broader in scope with more constructed-response questions and critical thinking, while SAT required more symbol manipulation and contained items that were less complex.

When asked about the facilities for the alignment workshop, panelists generally felt that they were suitable. Table 31 shows panelist responses to this question, by panel. Numbers in bold font represent the highest number of responses for each question, by panel. Six panelists in Panel 2 responded to these questions; therefore, only six responses are reported for this panel.

Table 31. Panelist Responses Regarding Adequacy of Facilities

	Panel 1 (n=8)				**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	3	**5**	0	0	**3**	**3**
Computers and equipment	0	2	**3**	**3**	0	0	2	**4**
Meals and breaks	0	0	0	**8**	0	0	1	**5**
Sleeping rooms	0	0	0	**8**	0	2	0	**4**

Across both panels, all the panelists felt that the meeting rooms and meals and breaks were either adequately suitable or very suitable. All but two of the panelists reported that the computers and equipment and the sleeping rooms were either adequately suitable or very suitable for this type of meeting; panelists who indicated that the equipment and sleeping rooms were only somewhat suitable did not provide any further comments.

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 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 mathematics assessments.

Summary of Overlap of Content Alignment

Table 32 shows the overlap of content alignment of each assessment to its own and the other assessment's 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 and SAT Framework at the Standard Level

NAEP Framework	**NAEP Items** (42 items)		**SAT Items (Form D)** (54 items)		**SAT Items (Form E)** (54 items)	
	Panel 1	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
	% of Total Hits		**% of Total Hits**		**% of Total Hits**	
1–Number properties and operations	22	22	23	24	22	22
2–Measurement	16	16	11	11	7	9
3–Geometry	16	16	20	21	17	19
4– Data analysis, statistics, and probability	11	12	10	10	13	12
5–Algebra	34	34	36	34	41	39
SAT Framework	**NAEP Items** (164 items)[17]		**SAT Items** (40 items)			
	Panel 1	**Panel 2**	**Panel 1**		**Panel 2**	
	% of Total Hits		**% of Total Hits**		**% of Total Hits**	
N–Number and operations	19	21	22		20	
A–Algebra and functions	33	31	35		37	
G–Geometry and measurement	27	27	34		34	
D–Data analysis, statistics, and probability	21	21	9		9	

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

NAEP items were found to assess all five NAEP standards (Standard 1, "Number properties and operations"; Standard 2, "Measurement"; Standard 3, "Geometry"; Standard 4, "Data analysis, statistics, and probability"; and Standard 5, "Algebra") and all of the four SAT standards (N, "Number and operations"; A, "Algebra and functions"; G, "Geometry and measurement"; and D, "Data analysis, statistics, and probability"). SAT items were found to assess all of the SAT standards and all NAEP standards.

With regard to alignment to the NAEP framework, both SAT items and NAEP short-version items had the highest percentage of their overall hits to "Algebra" (from 34% to 41% across both

[17] The percentages in this table indicate the distribution of total hits. It should be noted that, as shown in Table 21, 4% and 2% of the adjusted total hits for NAEP items were determined by panelists to be uncodable to any objective.

SAT forms and panels, and 34% across panels for NAEP). The lowest percentages of NAEP items in the short-version were found to align to “Data analysis, statistics, and probability” (11% and 12), while the lowest percentages in either SAT form were in Form E for “Measurement” (7% and 9%).

In relation to the SAT framework, both assessments had somewhat similar distributions of item alignments among the four SAT standards, with some differences in relative emphasis. Specifically, the SAT items had a greater proportion of items coded to “Geometry and measurement,” while NAEP had a greater proportion of items coded to “Data analysis, statistics, and probability.” For SAT items, “Algebra” received the highest percentages of item alignments (35% and 37%), followed closely by “Geometry and measurement” (34%). “Number and operations” received 22% and 20% of alignments. “Data analysis, statistics, and probability” had the lowest percentage of hits in the SAT assessment, with 9% of total hits.

Of the NAEP items that aligned to the SAT framework, the highest percentages of alignments, as in the SAT short-version items, were for “Algebra and functions” (31% and 33%), 27% of total hits were for “Geometry and measurement,” 21% of total hits were for “Data analysis, statistics, and probability,” and 19% and 21% of total hits were for “Number and operations.” Five NAEP items were found to be uncodable by the majority of panelists in each panel (5 or more of 8) and were not aligned to any SAT objective.

Overall, the NAEP items had hits to all of the SAT standards, with a slightly higher proportion of items coded to “Algebra and functions.” SAT short-version items had hits to all of the SAT standards, with a slightly higher proportion of items coded to “Algebra and functions” and “Geometry and measurement.”

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

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

NAEP Framework			**NAEP Items** (42 items)		**SAT Items (Form D)** (54 items)		**SAT Items (Form E)** (54 items)	
			Panel 1	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
Standards	**Goals**	**Objectives**	**% of Total Hits**		**% of Total Hits**		**% of Total Hits**	
1	1	1	0	0	0	0	0	0
	1.1	1.1.d	4	4	4	2	6	2
		1.1.f	2	2	0	0	0	0
		1.1.g	1	0	2	1	1	2
		1.1.i	0	0	0	1	0	0
	1.2	1.2.b	3	0	0	0	0	0
		1.2.c	0	0	0	0	0	0
		1.2.d	0	0	0	0	0	0
		1.2	1	1	1	0	0	0

			NAEP Items (42 items)		**SAT Items (Form D)** (54 items)		**SAT Items (Form E)** (54 items)	
NAEP Framework			**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
Standards	**Goals**	**Objectives**	**% of Total Hits**		**% of Total Hits**		**% of Total Hits**	
	1.3	1.3.a	0	1	0	0	0	0
		1.3.b	3	6	2	3	0	2
		1.3.c	2	2	0	0	0	0
		1.3.d	0	0	2	2	0	2
		1.3.f	0	2	2	5	4	4
		1.3	0	0	0	0	0	0
	1.4	1.4.c	0	1	3	4	6	5
		1.4.d	3	3	3	3	2	2
		1.4	0	0	0	0	0	0
	1.5	1.5.c	2	2	2	1	0	1
		1.5.d	0	0	1	0	1	2
		1.5.e	0	0	1	1	0	0
		1.5.f	0	0	0	0	0	2
		1.5	0	0	0	0	0	0
	1.6	1.6.a	0	0	1	0	0	0
		1.6.b	2	0	0	0	0	0
		1.6	0	0	0	0	0	0
2	2.1	2.1.b	0	0	0	0	0	0
		2.1.c	0	2	0	0	0	0
		2.1.d	0	0	5	6	4	4
		2.1.f	4	1	5	5	3	3
		2.1.h	0	0	1	0	0	1
		2.1.i	1	3	0	0	0	1
		2.1	0	0	0	0	0	0
	2.2	2.2.a	2	2	0	0	0	0
		2.2.b	2	1	0	0	0	0
		2.2.d	0	0	0	0	0	0
		2.2.e	0	0	0	0	0	0
		2.2.f	4	5	0	0	0	0
	2.3	2.3.a	0	0	0	0	0	0
		2.3.b	0	0	0	0	0	0
		2.3.c	2	2	0	0	0	0
		2.3.d	0	0	0	0	0	0
		2.3.e	0	0	0	0	0	0
		2.3.f	0	0	0	0	0	0

NAEP Framework			**NAEP Items** (42 items)		**SAT Items (Form D)** (54 items)		**SAT Items (Form E)** (54 items)	
			Panel 1	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
Standards	**Goals**	**Objectives**	**% of Total Hits**		**% of Total Hits**		**% of Total Hits**	
		2.3.g	0	0	0	0	0	0
3	3.1	3.1.c	0	0	0	0	0	0
		3.1.d	0	0	0	0	0	0
		3.1.e	0	0	1	3	0	0
		3.1.f	0	0	0	0	0	0
	3.2	3.2.a	0	0	0	0	1	1
		3.2.b	0	0	0	0	0	0
		3.2.c	0	0	0	0	0	0
		3.2.d	2	2	0	0	0	0
		3.2.e	0	0	0	0	0	0
		3.2.g	0	0	0	0	0	0
		3.2	0	0	0	0	0	0
	3.3	3.3.b	1	1	7	4	5	4
		3.3.c	0	0	0	0	1	0
		3.3.d	3	2	3	3	2	4
		3.3.e	1	0	0	0	1	2
		3.3.f	2	2	0	1	0	0
		3.3.g	0	0	2	1	1	1
		3.3.h	1	2	1	2	3	1
	3.4	3.4.a	2	3	3	4	4	5
		3.4.b	0	0	0	0	0	0
		3.4.c	2	2	0	0	0	0
		3.4.d	0	0	1	1	0	0
		3.4.e	0	0	0	0	0	0
		3.4.f	0	0	0	0	0	0
		3.4.g	0	0	0	0	0	0
		3.4.h	2	2	0	0	0	0
		3.4	0	0	0	0	0	0
	3.5	3.5.a	0	0	0	0	0	0
		3.5.b	0	0	0	0	0	0
		3.5.c	0	0	0	0	0	0
		3.5.d	0	0	0	0	0	0
		3.5.e	0	0	0	0	0	0
4	4	4	0	0	0	0	0	0
	4.1	4.1.a	3	4	5	4	4	3

			NAEP Items (42 items)		**SAT Items (Form D)** (54 items)		**SAT Items (Form E)** (54 items)	
NAEP Framework			**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**	**Panel 1**	**Panel 2**
Standards	**Goals**	**Objectives**	**% of Total Hits**		**% of Total Hits**		**% of Total Hits**	
		4.1.b	0	0	0	0	0	0
		4.1.c	0	0	0	0	0	0
		4.1.d	1	1	0	0	0	0
		4.1.e	0	0	0	0	0	0
		4.1.f	0	0	0	0	0	0
	4.2	4.2.a	0	0	2	2	3	4
		4.2.b	0	0	0	0	0	0
		4.2.c	0	0	0	0	2	0
		4.2.d	1	1	0	0	0	0
		4.2.e	0	0	0	0	0	0
		4.2.f	0	0	0	0	0	0
		4.2.g	0	0	0	0	0	0
	4.3	4.3.a	0	0	0	0	0	0
		4.3.b	0	0	0	0	0	0
		4.3.c	1	0	0	0	0	0
		4.3.d	0	0	0	0	0	0
		4.3.e	0	0	0	0	0	0
	4.4	4.4.a	0	0	0	0	0	0
		4.4.b	0	0	1	2	1	1
		4.4.c	1	2	0	0	0	0
		4.4.d	0	0	0	0	0	0
		4.4.e	0	0	2	2	1	2
		4.4.h	1	2	0	0	0	0
		4.4.i	0	0	0	0	0	0
		4.4.j	0	0	0	0	0	0
		4.4.k	0	0	0	0	0	0
	4.5	4.5.a	1	2	0	0	0	0
		4.5.b	0	0	0	0	0	0
		4.5.c	0	0	0	0	0	0
		4.5.d	0	0	0	0	0	0
		4.5.e	0	0	0	0	0	0

			NAEP Items (42 items)		SAT Items (Form D) (54 items)		SAT Items (Form E) (54 items)	
NAEP Framework			Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
Standards	Goals	Objectives	% of Total Hits		% of Total Hits		% of Total Hits	
5	5	5	0	0	0	0	0	0
	5.1	5.1.a	2	4	4	5	3	3
		5.1.b	4	2	0	0	0	0
		5.1.e	2	5	0	0	3	4
		5.1.g	0	0	0	0	0	0
		5.1.h	3	0	0	0	0	1
		5.1.i	2	0	0	0	0	0
		5.1.j	0	0	0	0	0	0
	5.2	5.2.a	2	4	4	2	5	6
		5.2.b	1	0	0	0	1	0
		5.2.d	1	1	0	0	0	0
		5.2.e	0	0	0	0	0	0
		5.2.f	2	2	0	0	0	0
		5.2.g	0	0	0	0	0	0
		5.2.h	0	0	0	0	0	0
	5.3	5.3.b	0	1	3	2	4	4
		5.3.c	3	2	2	3	1	2
		5.3.d	3	2	0	3	6	4
		5.3.e	4	3	0	5	0	0
		5.3.f	2	2	3	4	4	2
		5.3.g	0	0	0	1	0	0
		5.3.h	0	0	0	0	0	0
	5.4	5.4.a	2	0	6	2	6	6
		5.4.c	0	0	2	0	4	3
		5.4.d	0	0	6	3	3	3
		5.4.e	0	0	0	0	0	0
		5.4.f	0	0	4	3	0	0
		5.4.g	0	0	0	0	0	0
	5.5	5.5.a	0	5	0	0	0	0
		5.5.b	2	0	0	0	0	0
		5.5.c	0	0	0	0	0	0
		5.5	0	0	0	0	0	0

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

As shown in Table 33, NAEP and SAT items were evenly distributed across the NAEP objectives, with broader overall coverage by the NAEP items (72 objectives hit, compared to 51 for SAT). Most of the hits from the NAEP short-version items ranged between 1% and 4% of total hits per objective; only three NAEP objectives received more than 4% of the total hits: Objective 1.3.b (“Perform arithmetic operations with real numbers, including common irrational numbers”), Objective 2.2.f (“Construct or solve problems involving scale drawings”), and Objective 5.1.e (“Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or *trigonometric functions from tables, graphs, or equations”). Similarly, most of the hits from SAT items ranged between 1% and 6% of total hits per objective. Only one NAEP objective received more than 6% of the total hits: Objective 3.3.b, “Apply geometric properties and relationships to solve problems in two and three dimensions.”

In order to elucidate the differences in emphasis between the two assessments, objectives have been identified that had aligned items in one assessment but only minimal coverage in the other assessment. The following NAEP objectives had at least 1% of total hits from the NAEP items in the short-version (42 items) in both panels but less than 1% of total hits from the two forms of SAT (54 items per form) in any one panel:

- 1.1.f, “Represent or interpret expressions involving very large or very small numbers in scientific notation”
- 1.3.c, “Perform arithmetic operations with expressions involving absolute value”
- 2.1.i, “Solve problems involving rates such as speed, density, population density, or flow rates”
- 2.2.a, “Recognize that geometric measurements (length, area, perimeter, and volume) depend on the choice of a unit, and apply such units in expressions, equations, and problem solutions”
- 2.2.b, “Solve problems involving conversions within or between measurement systems, given the relationship between the units”
- 2.2.f, “Construct or solve problems involving scale drawings”
- 2.3.c, “Use the definitions of sine, cosine, and tangent as ratios of sides in a right triangle to solve problems about length of sides and measure of angles”
- 3.2.d, “Identify transformations, combinations, or subdivisions of shapes that preserve the area of two-dimensional figures or the volume of three-dimensional figures”
- 3.3.f, “Analyze properties or relationships of triangles, quadrilaterals, and other polygonal plane figures”
- 3.4.c, “Describe or identify conic sections and other cross sections of solids”
- 3.4.h, “*Represent situations and solve problems involving polar coordinates”
- 4.1.d, “Given a graphical or tabular representation of a set of data, determine whether information is represented effectively and appropriately”
- 4.2.d, “Compare data sets using summary statistics (mean, median, mode, range, interquartile range, or standard deviation) describing the same characteristic for two different populations or subsets of the same population”
- 4.4.c, “Given the results of an experiment or simulation, estimate the probability of simple or compound events in familiar or unfamiliar contexts”
- 4.4.h, “Determine the probability of independent and dependent events”
- 4.5.a, “Identify misleading uses of data in real-world settings and critique different ways of presenting and using information”

- 5.1.b, “Express linear and exponential functions in recursive and explicit form given a table, verbal description, or some terms of a sequence”
- 5.2.d, “Perform or interpret transformations on the graphs of linear, quadratic, exponential, and *trigonometric functions”
- 5.2.f, “Given a real-world situation, determine if a linear, quadratic, rational, exponential, logarithmic, or *trigonometric function fits the situation”

Conversely, SAT items were mapped by both panels to 14 NAEP objectives for which the NAEP short-version sample set of items had less than 1% of total hits from both panels.

The following NAEP objectives had at least 1% of total hits from SAT items on at least one form by both panels and less than 1% of hits from the NAEP short-version sample set from both panels:

- 1.3.d, “Describe the effect of multiplying and dividing by numbers including the effect of multiplying or dividing a real number by: Zero, or A number less than zero, or A number between zero and one, or One, or A number greater than one”
- 1.5.d, “Use divisibility or remainders in problem settings”
- 1.5.e, “Apply basic properties of operations, including conventions about the order of operations”
- 2.1.d, “Solve problems of angle measure, including those involving triangles or other polygons or parallel lines cut by a transversal”
- 3.1.e, “Use two-dimensional representations of three-dimensional objects to visualize and solve problems”
- 3.2.a, “Recognize or identify types of symmetries (e.g., point, line, rotational, self-congruence) of two- and three-dimensional figures”
- 3.3.g, “Analyze properties and relationships of parallel, perpendicular, or intersecting lines including the angle relationships that arise in these cases”
- 3.4.d, “Represent two-dimensional figures algebraically using coordinates and/or equations”
- 4.2.a, “Calculate, interpret, or use summary statistics for distributions of data including measures of typical value (mean, median), position (quartiles, percentiles), and spread (range, interquartile range, variance, and standard deviation)”
- 4.4.b, “Determine the theoretical probability of simple and compound events in familiar or unfamiliar contexts”
- 4.4.e, “Determine the number of ways an event can occur using tree diagrams, formulas for combinations and permutations, or other counting techniques”
- 5.4.c, “Analyze situations, develop mathematical models, or solve problems using linear, quadratic, exponential, or logarithmic equations or inequalities symbolically or graphically”
- 5.4.d, “Solve (symbolically or graphically) a system of equations or inequalities and recognize the relationship between the analytical solution and graphical solution”
- 5.4.f, “Solve an equation or formula involving several variables for one variable in terms of the others”

The preceding lists can be used to examine differences in breadth and emphasis for the two assessments across all standards and objectives. For example, for all standards except “Geometry,” each assessment had hits to a similar number of objectives for which the other assessment had little coverage (less than 1% of total hits).

Differences between the assessments were also observed at the goal level.

The goals for which the NAEP items appear to have greater emphasis are as follows:

- 1.2, “Estimation”
- 2.2, “Systems of measurement”
- 5.1, “Patterns, relations, and functions”

The goals for which the SAT items appear to have greater emphasis are as follows:

- 2.1, “Measuring physical attributes”
- 3.1, “Dimension and shape”
- 3.3, “Relationships between geometric figures”
- 4.2, “Characteristics of data sets”
- 5.4, “Equations and inequalities”

Despite some differences, there were some commonalities across the two tests at the objective level. Both had at least 1% of hits to 16 objectives:

- 1.1.d, “Represent, interpret, or compare expressions for real numbers, including expressions using exponents and logarithms”
- 1.3.b, “Perform arithmetic operations with real numbers, including common irrational numbers”
- 1.4.d, “Solve multistep problems involving percentages, including compound percentages”
- 1.5.c, “Solve problems using factors, multiples, or prime factorization”
- 2.1.f, “Solve problems involving perimeter or area of plane figures such as polygons, circles, or composite figures”
- 3.3.b, “Apply geometric properties and relationships to solve problems in two and three dimensions”
- 3.3.d, “Use the Pythagorean theorem to solve problems in two- or three-dimensional situations”
- 3.3.h, “Analyze properties of circles and the intersections of lines and circles (inscribed angles, central angles, tangents, secants, and chords)”
- 3.4.a, “Solve problems involving the coordinate plane such as the distance between two points, the midpoint of a segment, or slopes of perpendicular or parallel lines”
- 4.1.a, “Read or interpret graphical or tabular representations of data”
- 5.1.a, “Recognize, describe, or extend numerical patterns, including arithmetic and geometric progressions”
- 5.1.e, “Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or *trigonometric functions from tables, graphs, or equations”

- 5.2.a, “Create and translate between different representations of algebraic expressions, equations, and inequalities (e.g., linear, quadratic, exponential, or *trigonometric) using symbols, graphs, tables, diagrams, or written descriptions”
- 5.3.c, “Perform basic operations, using appropriate tools, on algebraic expressions including polynomial and rational expressions”
- 5.3.d, “Write equivalent forms of algebraic expressions, equations, or inequalities to represent and explain mathematical relationships”
- 5.3.f, “Use function notation to evaluate a function at a specified point in its domain and combine functions by addition, subtraction, multiplication, division, and composition”

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

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

SAT Framework		**NAEP Items** (164 items)		**SAT Items** (40 items)	
		Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
Standards	**Objectives**	**% of Total Hits**		**% of Total Hits**	
N–Number and operations	N	1	0	0	0
	N.1	2	2	5	3
	N.2	3	2	1	0
	N.3	8	10	3	3
	N.4	2	3	5	6
	N.5	0	1	1	0
	N.6	1	1	2	2
	N.7	2	3	6	5
A–Algebra and functions	A	4	5	0	0
	A.1	4	5	4	6
	A.2	6	6	11	9
	A.3	2	1	4	3
	A.4	1	1	4	5
	A.5	1	1	2	2
	A.6	0	0	0	0
	A.7	1	1	2	2
	A.8	2	2	2	2
	A.9	0	0	0	2
	A.10	12	10	5	6

SAT Framework		**NAEP Items** (164 items)		**SAT Items** (40 items)	
		Panel 1	**Panel 2**	**Panel 1**	**Panel 2**
Standards	**Objectives**	**% of Total Hits**		**% of Total Hits**	
G–Geometry and measurement	G	5	4	0	0
	G.1	2	2	3	2
	G.2	1	1	3	3
	G.3	1	1	5	5
	G.4	4	5	6	4
	G.5	2	2	6	7
	G.6	3	1	3	2
	G.7	2	3	0	0
	G.8	4	4	2	3
	G.9	3	5	6	6
D–Data analysis, statistics, and probability	D	5	1	0	0
	D.1	8	9	4	4
	D.2	2	5	2	3
	D.3	5	6	3	2

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

In relation to the SAT framework, the distribution of item alignments to objectives appeared relatively similar on both tests, with almost all objectives receiving 1% or more total hits in both panels. A few objectives received more than 5% of total hits on either or both tests.

SAT objectives that received 5% or more of total hits for the alignment of the NAEP items from at least one panel are listed below:

- N.3, “Arithmetic word problems”
- A.2, “Algebraic representations, translations, and algebraic word problems”
- A.10, “Basic concepts of algebraic functions”
- D.1, “Data interpretation”
- D.3, “Probability”

No objective for “Geometry” received more than 5% of NAEP total hits; however, every objective in this standard did receive at least 1% from each panel.

SAT objectives that received 5% or more of total hits for the alignment of the SAT items from at least one panel are listed below:

- N.4, “Sequences and series”
- N.7, “Logic/logical reasoning”
- A.1, “Operations with real numbers”
- A.2, “Algebraic representations, translations, and algebraic word problems”
- A.10, “Basic concepts of algebraic functions”
- G.4, “Special triangles”
- G.5, “Circles”
- G.9, “Coordinate geometry”

No objective for “Data analysis, statistics, and probability” received more than 5% of SAT total hits; however, every objective in this standard did receive at least 2% from each panel.

For each SAT standard, some NAEP items were aligned at the standard level but not to any objective within the standard. This was not the case with the SAT items, which were all coded at the objective level.

There were two objectives in the SAT framework that did not receive at least 1% of total hits in either the NAEP items or the SAT items: Objective A.6, “Radical equations,” and Objective A.9, “Direct and inverse variation.”

As shown in Table 34, within the SAT framework, NAEP items had more overall coverage of the objectives within “Data analysis, statistics, and probability” (21% of total hits for both panels to the three objectives, compared to 9% of total SAT hits for both panels to the three objectives). SAT items had similar (only slightly greater) coverage of objectives for the other three standards: “Number and operations” had 20% and 22% of total SAT hits compared to 19% and 21% of total NAEP hits; “Algebra and functions” had 35% and 37% of total SAT hits, compared to 31% and 33% of total NAEP hits; and “Geometry and measurement” had 34% of total SAT hits for both panels, compared to 27% of total NAEP hits for both panels.

Across all standards, the two tests had at least 1% of total hits from both panels to 24 of the 29 objectives, with similar percentages (the same or within 2%) of hits to 13 of those objectives:

- N.5, “Sets”
- N.6, “Counting problems”
- A.1, “Operations with real numbers”
- A.3, “Linear equations and inequalities”
- A.5, “Rational equations and inequalities”
- A.7, “Quadratics”
- A.8, “Manipulation with integer and rational exponents and using rules for exponents”
- G.1, “Points and lines in the plane”
- G.2, “Angles in the plane”
- G.4, “Special triangles”
- G.6, “Polygons”
- G.8, “Geometric perception”
- D.2, “Statistics”

Although the differences between the two tests appear less pronounced in relation to the SAT framework than in relation to the NAEP framework, this may be partly due to the more general language of the SAT objectives compared to the more specific NAEP objectives. For example, SAT Objective A.10 (“Basic concepts of functions”) received a high percentage of total hits (10% and 12% for the NAEP items and 5% and 6% for the SAT items). The items coded to that objective would likely have been coded to several objectives within NAEP Goal 5.1 (“Patterns, relations, and functions”). In one such case, two NAEP items that were aligned to SAT Objective A.10 and also included in the sample for Sub-Study 1 were aligned to two different NAEP objectives in Goal 5.1 (5.1.b and 5.1.i).

Categorical Concurrence

For alignment to the NAEP framework, the NAEP items used in the short-version (42 items) were found to meet the typical WAT threshold value of at least six items for categorical concurrence for four of the five standards. In “Data analysis, statistics, and probability,” the WAT values were 4.56 and 4.88 mean hits. The SAT items were found to meet the criterion for three of the five standards for Form D and four of the five standards for Form E. In Form D, the criterion was on the borderline of being met for “Measurement,” with 5.62 and 6 mean hits, and was not met for “Data analysis, statistics, and probability,” with 5.5 mean hits for both panels. In Form E, the criterion was not met for “Measurement,” with 3.88 and 4.62 mean hits.

For alignment to the SAT framework, both the NAEP items (164 items) and the SAT items were found to meet the typical WAT threshold value of at least six items for categorical concurrence for all five standards.

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

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

For alignment to the NAEP framework, the NAEP items were found to meet depth-of-knowledge consistency in all standards, except for 46% recorded for “Data analysis, statistics, and probability” by Panel 1. 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 the NAEP standards, except for 42% recorded for “Measurement” for Form E by Panel 2.

For alignment to the SAT framework, DOK was analyzed as range of depth of knowledge. Items that aligned to the SAT framework were coded at DOK Levels 1, 2, or 3. Most of the NAEP items were coded at DOK Level 2 (47% and 61% of total mean hits), and a substantial number were coded at DOK Level 1 (46% and 27%). The remaining items were coded to DOK Level 3 (5% and 5%). Similarly to the NAEP items, most of the SAT short-version items were coded at DOK Level 2 (64% and 77%), and a substantial number were coded at DOK Level 1 (30% and 17%). The remaining items were coded to DOK Level 3 (7% and 12%).

Range-of-Knowledge Correspondence

For alignment to the NAEP framework, the NAEP short-version set of items did not meet the criteria for range of knowledge for any standard (50% or more of objectives hit). No NAEP standard had more than 42% of its objectives hit by items in the short-version; “Data analysis, statistics, and probability” had the most restricted range of knowledge, with only 12% and 14% of its objectives hit. This result likely reflects the large number of objectives (130 objectives) relative to the number of items in the short form (42 items) used in this study. For alignment of the SAT items (108 items) to the NAEP framework, range of knowledge was weakly met for “Number properties and operations,” and was not met for the other four standards.

For alignment to the SAT framework, the NAEP items (164 items) had a range of knowledge above the 50% criterion for all four standards. For the SAT items (40 items), the range of knowledge criterion was met for all four of the SAT standards.

Balance of Representation

Both the NAEP short-version items and the SAT items met the criteria for balance of representation for all five standards in the NAEP framework.

In relation to the SAT framework, the NAEP items met the criteria for balance of representation for "Geometry and measurement" and "Data analysis, statistics, and probability" and weakly met the criteria for "Number and operations" and "Algebra and functions." The SAT items met the criteria for balance of representation for all four standards.

Overall Conclusions

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

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

At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining Geometry and measurement into one standard. Each framework contains both general and specific objectives, although the SAT objectives, which are presented as content topics without indication of the cognitive level at which that content would be assessed, may be interpreted as more general than the NAEP objectives.

Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are all addressed in somewhat similar proportions.

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

The greatest commonality between the two tests is in their relatively similar emphasis at the standard level. This is evident in the distribution of percentages of total hits from both tests matched to each set of standards. Although there are some differences of emphasis, such as the full NAEP pool's greater proportion of alignment to SAT "Data analysis, statistics, and probability," and the SAT short-version's greater proportion of alignment to SAT "Geometry and measurement," the proportions of alignments to "Algebra and functions" and "Number and operations" are very close. There is also considerable overlap among some specific skills, with both tests addressing many of the same NAEP "Number properties and operations" objectives (such as 1.1.d, 1.4.d., 3.3.d, 4.1.a, and 5.3.c) and SAT objectives (such as N.6, A.10, D.2, and G.1). Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), it is clear that both tests emphasize a number of the same or closely related skills. These include properties, equivalence, and operations on rational numbers (included in NAEP Goals 1.1 and 1.3 and included in SAT Objective N.2) and properties of two-dimensional shapes (included in NAEP Goals 3.1 and 3.3 and included in SAT Objective G.6).

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 mathematics subdomains are missing or not aligned?

While there is considerable overlap between the two assessments, primarily in the intersection of the NAEP “Algebra” and SAT “Algebra and functions” standards, there are notable differences as well. The SAT items had a somewhat limited range of coverage of the NAEP standards “Measurement,” “Geometry,” and “Data analysis, statistics, and probability,” with several goals receiving few hits. However, even given the minimal coverage of some of the goals within each NAEP standard by SAT items, almost all NAEP items found a match in the SAT framework. The language of the objectives in the SAT framework is sufficiently broad to encompass the range of the NAEP items. For example, SAT Objective A.10, “Basic concepts of algebraic functions,” may accommodate most of the items aligning to the seven objectives within NAEP Goal 5.1, “Patterns, relations, and functions.” In fact, several items on the NAEP short form that were coded to objectives in Goal 5.1 in the NAEP framework were matched to SAT Objective A.10 in the SAT framework. Finally, some NAEP items were found to be uncodable to the SAT objectives. These items assessed skills not present in the SAT framework.

The two tests are also similar in the average DOK levels of items. However, while most items in both tests were found to be at DOK Level 2, NAEP items had a wider range of DOK than did SAT items, with more items coded to Levels 1 and 3. The Level 3 NAEP items often involved application of concepts through short or extended constructed-response items. Both tests also met depth-of-knowledge consistency overall (with each not meeting this criterion for only one standard as rated by one panel).

Overall, despite differences in alignment at the detailed specific objective level, differences in emphasis at the standard level, and a small difference in ranges of depth of knowledge, there is considerable overlap of content and complexity between the two assessments.

VI. Discussion and Recommendations on Study Design

This alignment study involved the implementation of a study design custom-developed by Dr. Webb. Given the relatively early stage of the field of assessment-to-assessment alignment, and at the request of the Governing Board, this section includes some considerations and recommendations related to implementation of the study design during the pilot study and the operational studies (NAEP–ACCUPLACER reading and mathematics, and NAEP–SAT reading and mathematics). Process recommendations from the pilot study are included in Section II of this report and in the Pilot Study Report. In addition, some of the recommendations from the Pilot Study Report are restated here, as they relate to the overall study design. Except where specifically related to mathematics or 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 mathematics study, WestEd received from the Governing Board and the College Board very different framework documents with different levels of specificity of content for NAEP and SAT. Among the most substantive differences was the NAEP framework’s inclusion of language describing how students would apply the knowledge and skills, while the SAT framework focused on content topics.

In interpreting the study results, it is also important to consider that panelists were selecting from 130 NAEP objectives across five standards and 29 broader SAT objectives across four standards. Each framework had some internal overlap among the content of the objectives; indeed, in the domain of mathematics there is overlap among content, skills, and approaches necessary to solve problems. The large number of objectives in the NAEP framework affected a number of areas of the study. First, it increased coding time, since there were more objectives to analyze. It also increased the likelihood that discrepancies in coding items to objectives would need to be discussed. Finally, it increased the likelihood that some objectives would not be matched to one or more items. Conversely, with fewer, broader objectives, there is an increased likelihood that the SAT framework would receive hits from NAEP items across the range of objectives.

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 mathematics. In particular, there was some discussion among panelists and facilitators about the difference between problems requiring multiple computation steps (that is, the application of the same or similar concept multiple times) and problems involving multiple concepts, and whether the former should be considered as DOK Level 1 or Level 2 for grade 12 students. In other cases, the clarity of the wording of the DOK level descriptions prompted discussion about appropriate interpretation.

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

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

Order of Sub-Studies

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

Placement of Correct Answers in Item Booklets

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

Cross-Panel Adjudication

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

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

Data Analysis

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

Another issue related to data analysis that required follow-up discussion was how to use the replicate panel data. The design document indicates that the results could be aggregated or averaged once it was established that the panels were indeed replicate. However, the WAT system is not currently programmed to combine studies in this way. Following discussions with the Governing Board, Dr. Webb, and ACT, it was decided to report both panels' results separately in order to show areas where the replicate panels produced discrepant results, which may in itself be an interesting finding regarding alignment.

Other Factors That May Affect Alignment

The alignment methodology used in this study captures the degrees of alignment between the assessments and their respective 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, in reading specifically, although items may be aligned to the same objectives, the amount and level of reading (not just genre) may be an important difference between the two assessments in how they assess reading and 12th grade preparedness. Similarly, it is possible that there are other preparedness-related differences between the content of the assessments—related, for instance, to the variety of item types on NAEP (i.e., multiple choice, short constructed response, extended constructed response) in comparison with the variety of types on SAT (i.e., multiple choice, student-produced response)—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 this report and WestEd's comprehensive report on the NAEP–ACCUPLACER mathematics study, this presented a challenge to keeping the mathematics panels to the allotted schedule. Panelist feedback on the long coding sessions confirms that, for mathematics, either a reduction in workload or an extension of the length of the workshop would be desirable. Therefore, monitoring overall workload should be an explicit objective of the study design.

Panelist Experience

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

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

VII. References

The College Board. (2007). *SAT® Skills Insight™ mathematics 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). *Mathematics Framework for the 2009 National Assessment of Educational Progress.* Developed for the National Assessment Governing Board in support of Contract No. ED-00-CO-0115, U.S. Department of Education, by the Council of Chief State School Officers, with subcontracts to the Council of Basic Education and the Association of State Supervisors of Mathematics and Grade 12 preparedness objectives developed under contract with Achieve, Inc.

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

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

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

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

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