The Alignment of the NAEP Grade 12 *Reading* Assessment and the WorkKeys *Reading for Information* Assessment

October 2010

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

Table of Contents

List of Tables

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 twelfth-grade students in reading and mathematics for postsecondary education and training.

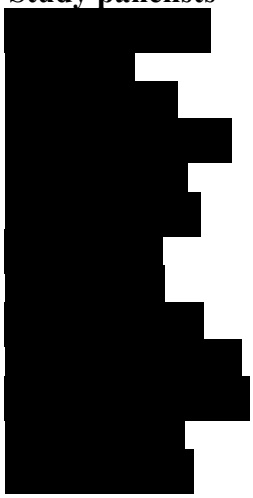
The findings and conclusions presented in this research report, by themselves, do not support statements about twelfth-grade student preparedness in relation to NAEP reading and mathematics results. Readers should not use the findings and conclusions in this report to draw conclusions or make inferences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

Acknowledgements

This study was funded by the National Assessment Governing Board under Contract ED-06-CO-0098 and managed by staff of the Workforce Development Division of ACT, Inc.

Study co-facilitators
Cynthia Jacobson
John Fortier

Study panelists



ACT, Inc. staff
Oliver Cummings
Jennifer Horn-Frasier
Edythe Thompson

Executive Summary

The National Assessment of Educational Progress (NAEP) is a nationally representative testing program that measures student academic achievement. In 2004, a recommendation was made that the NAEP be used to report on the preparedness of the nation's twelfth-graders for postsecondary endeavors including college, training for employment, and entrance into the military. Therefore, the National Assessment Governing Board (NAGB) sought to study, using a rigorous evaluation process, the extent to which the NAEP for reading and mathematics might be used as an indicator of preparedness for training for occupations. NAGB has established a research program to explore this issue.

This report describes the result of one study in this research program, the alignment between the NAEP Grade 12 *Reading* assessment and ACT, Inc.'s WorkKeys *Reading for Information* assessment. The WorkKeys assessment is a widely recognized standardized test related to the workplace, and that is why it was selected for this study. The alignment study was conducted over the course of a week in January 2010 at ACT's national headquarters in Iowa City, IA, using two concurrent, replicate panels of reading content experts from across the United States.

The alignment study was designed to follow methodology developed by Dr. Norman Webb; the study design document is included in Appendix A. Webb's methodology has been used many times to study the alignment of tests to the standards on which they are based. This particular study is a special application of Webb's methodology; it is an assessment-to-assessment alignment study, rather than an assessment-to-standards alignment study. The methodology makes use of two concurrent, replicate panels of experts. The two panels were combined for training to ensure that all participants received the same information, and they worked separately for most of the rest of the tasks. The two facilitators communicated throughout each day's work and also in the evenings to identify areas their respective panels should discuss further and to plan any necessary adjustments to the procedures.

Although the documents from which the content representation used in the study was derived for the two assessments do not necessarily refer to them as "standards," this term will be used in this report for the purpose of simplicity. The documents that served as the standards are in Appendix E.

Webb has defined four depth of knowledge (DOK) levels (Level 1 to Level 4), which range from simple, fact-oriented knowledge and skills to deep knowledge and higher-order thinking skills. Reading assessment materials at DOK Level 1 typically involve basic comprehension or slight paraphrasing. Those at DOK Level 2 involve both comprehension and subsequent processing of text, such as summarizing, comparing, or identifying as fact or opinion. At DOK Level 3, reading assessment materials focus on deep knowledge and involve activities such as reasoning, planning, analyzing, providing support for thinking, and summarizing information from multiple sources. DOK Level 4 reading assessment materials require higher-order thinking and deep knowledge, and they typically require an extended period of time to complete a task, which often includes applying information from one source to a new task in such ways as analyzing information from multiple texts or explaining alternative perspectives across a variety of sources. See Appendix D of this report for the full description of these levels as used with the panelists for this study.

The two concurrent, replicate panels determined the DOK level of each NAEP and WorkKeys test standard and test item. The study methodology required the two panels to achieve consensus on the DOK levels for the standards, so the two groups were combined for an adjudication process to accomplish this. The methodology did not require such consensus for the DOK levels of test items; therefore, the two panels worked independently on the DOK levels of the items used in the study.

The DOK results may be summarized as follows:

- The range of DOK levels assigned to the NAEP standards was 1 – 4, and the average DOK level of the NAEP standards was 2.49.
- The range of DOK levels assigned to the WorkKeys standards was 1 – 3, and the average DOK level of the WorkKeys standards was 1.92.
- The range of DOK levels assigned to the NAEP items was 1 – 3, and the average DOK level for all NAEP items was 2.15
- The range of DOK levels assigned to the WorkKeys items was 1 – 2, and the average DOK level for all WorkKeys items was 1.54

Table ES1 shows key features of the two assessments, as delineated by the blueprint analysis and this study. Some of these features have an impact on the DOK level results.

Table ES1: Key features of the NAEP and WorkKeys assessments

Assessment Feature	NAEP Grade 12 *Reading* Assessment	WorkKeys *Reading for Information* Assessment
Item pool	All 131 items of the 2009 NAEP Grade 12 *Reading* item pool were used for this study.	A pool of 60 items drawn from the operational WorkKeys *Reading for Information* item pool of hundreds of items was used for this study.
Types of reading passages	3 of 15 documents used for this study had a workplace context; 1 was consumer oriented. • 30% literary nonfiction, fiction, or poetry • 31% informational expository • 27% argumentative/persuasive • 12% procedural	All 28 WorkKeys documents used for this study had a workplace context. • 32% policy • 35% instructions • 18% legal document • 15% information
Difficulty of reading passages	The difficulty of all reading passages is grade-12 appropriate.	The difficulty of reading passages ranges from grade 6 to postsecondary.
Types of items/Average DOK level	• 58% multiple choice / 1.74 • 32% short constructed response / 2.64 • 10% extended constructed response / 2.92	• 100% multiple choice / 1.54
Standards on which items are based / Average DOK level	1) Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension. / 1.50 2) Integrate/Interpret: Make complex inferences within and across texts. / 2.71 3) Critique/Evaluate: Consider text(s) critically. / 3.10	3) Individuals read short, simple, and clearly stated materials to find out what should be done. / 1.20 4) Individuals read straightforward information that contains a number of details. When following procedures, they must think about changing conditions that affect what should be done. / 1.75 5) Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions. / 1.83 6) Individuals read elaborate procedures, complicated information, and legal regulations, all of which contain difficult words, jargon, and technical terms. Most information is not clearly stated. / 2.43 7) Individuals read very complex information which includes a lot of details and complicated concepts. Unusual jargon and technical terms are used but not defined. Writing often lacks clarity and direction. Individuals must draw conclusions from some parts of the reading and apply them to other parts. / 2.33

In addition to assigning DOK levels to each test standard and test item, each panel completed the following sub-studies:

- Sub-Study 1: Map the NAEP items to the NAEP standards
- Sub-Study 2: Map the WorkKeys items to the NAEP standards

- Sub-Study 3: Map the NAEP items to the WorkKeys standards
- Sub-Study 4: Map the WorkKeys items to the WorkKeys standards

Throughout these four sub-studies, the two panels maintained a high level of interrater agreement, suggesting that it is appropriate to have confidence in the outcomes of the study.

Across the four sub-studies, the NAEP and WorkKeys test items were analyzed for their alignment with the three NAEP standards and the five WorkKeys standards according to four alignment criteria. This produced 64 points for which the degree of alignment was evaluated, using labels of Yes (alignment), Weak, and No (not aligned). The two concurrent panels reached the same conclusions for 51 of these points, and similar conclusions for another 11 of these points. There were just two points for which the two panels reached opposite conclusions (yes, aligned versus no, not aligned), despite following the prescribed adjudication processes. Thus, the replicate panels produced generally consistent judgments about the alignment of the tests. Results of each sub-study are given in detail in the body of this report.

In general, study results showed that the NAEP assessment covers a broad range of content across literary and informational reading, and students are asked to demonstrate the cognitive behaviors and skills of locating/recalling, integrating/interpreting, and critiquing/evaluating. On the other hand, the WorkKeys assessment covers a narrower range of content that focuses on reading procedural and policy/informational workplace documents, and examinees are asked to apply the content to workplace situations in which they must demonstrate skills such as determining next steps, following procedures, applying information to a situation not specifically described, or drawing conclusions and applying them to new situations. Skills measured by both assessments include identifying main ideas, details, and definitions; determining the correct meaning of a word based on context; explaining the rationale of a document; and identifying implied details.

The standards for the WorkKeys *Reading for Information* test are organized through five skill levels, moving from least to most complex. All texts are workplace communications found in identified career clusters, and they range from short, direct passages to longer, denser, more difficult materials (see Table ES1 for more details about the types of WorkKeys reading passages included in the pool that was analyzed for this study). WorkKeys texts exhibit great variability in reading level, clarity, and quality, as authentic workplace documents do; the reading difficulty of the passages ranges from sixth-grade to postsecondary level. WorkKeys items assess reading skills expected and needed for success in employment and in workforce training.

The WorkKeys items that aligned to the NAEP standards were related to locating and recalling information, causal relations, connecting ideas, drawing conclusions, providing supporting information, and determining word meaning in context. The WorkKeys items do not assess content described in the NAEP standards that is related specifically to literary reading passages, and neither do they assess NAEP content that involves critiquing or evaluating reading passages.

The NAEP standards are organized by cognitive target — from locating and recalling to integrating and interpreting to critiquing and evaluating — and type of text — fiction, poetry, literary nonfiction, informational expository, argumentative/persuasive, and procedural. Texts used on the NAEP assessment are well written and at a twelfth-grade level of difficulty. Thirty percent of the texts are literary and 70 percent are informational (see Table ES1 for more details

about the types of NAEP reading passages included in the pool that was analyzed for this study). Three of the 15 NAEP documents used for this study had an explicit workplace context, one document was oriented toward the consumer, and the remaining documents did not have an explicit workplace context. The NAEP items assess examinees' cognitive skills as applied to literary and informational texts.

Skills specified in the WorkKeys standards that are measured by the NAEP items include identifying main ideas, determining word meaning from context, explaining the rationale behind a text, and identifying implied details. Areas of the WorkKeys standards that are not assessed by the NAEP items are related to understanding, following, and applying instructions; determining and applying general principles contained in workplace documents and applying them to similar and new situations; and to the decoding of workplace jargon.

Throughout the study, which was very demanding for the participants, a great deal of qualitative feedback was elicited from the panelists. In general, this feedback indicated that the panelists felt comfortable with the process and positive about the experience. In addition, they felt that, while there is significant overlap between the content represented by the two tests, there are also important differences. One panelist summarized the differences this way: "Students who take NAEP are expected to meet cognitive targets on both literary and non-fiction texts. They are expected to have knowledge and skills that will lead to success in comprehending these materials. Students who take WorkKeys are expected to be able to apply, fairly immediately, what they learn from 'practical' texts such as rules, instructions, legal texts, etc."

Introduction

Purpose and the Governing Board's Approach to Preparedness

One important goal of K – 12 education is to prepare students for post-high school activities — postsecondary education, the military, or the workplace. Traditionally, the focus of standardized testing conducted at the end of high school has been on academic achievement or aptitude rather than on work-related skills.

The congressionally authorized National Assessment of Educational Progress (NAEP) is the only continuing source of comparable national and state data available to the public on the achievement of students at grades 4, 8, and 12 in core subjects. The National Assessment Governing Board (NAGB) oversees and sets policy for the NAEP. The NAEP and the Governing Board are authorized under the National Assessment of Educational Progress Authorization Act (P.L.107-279).

Among the Board's responsibilities is "to improve the form, content, use, and reporting of [NAEP results]." Toward this end, the Governing Board established a national commission to make recommendations to improve the assessment and reporting of NAEP at the twelfth grade. The commission issued its report in March of 2004. The commission noted the importance of maintaining the NAEP at the twelfth grade as a measure of the "output" of K – 12 education in the United States and as an indicator of the nation's human capital potential. The commission recommended that the Grade 12 NAEP be redesigned to report on the academic preparedness of twelfth-grade students in reading and mathematics for entry-level college credit coursework and for training for occupations. The commission concluded that having such information is essential for the economic well-being and security of the United States and that the NAEP is uniquely positioned to provide such information

As the Governing Board has been developing ways to implement the commission's recommendations, there has been a wider recognition — among federal and state policymakers, educators, and the business community — of the importance of a rigorous high school program that results in meaningful high school diplomas and prepares students for college and for training for good jobs.

The Governing Board has planned a program of research, consisting of 18 to 20 studies, to support the validity of statements about twelfth-grade student preparedness that would be made in NAEP reports, beginning with the 2009 assessments in twelfth-grade reading and mathematics. Included in the program of research are content alignment studies, to examine the degree of overlap of the domains measured by NAEP and a relevant assessment related to preparedness for college or job training.

The research described in this report addresses the alignment between the content of the NAEP Grade 12 Reading assessments as administered in 2009 and the content of the WorkKeys *Reading for Information* test. The WorkKeys assessment was selected because it is a widely recognized

standardized test related to the workplace. The Governing Board will use data resulting from this study, along with the results from other studies, to help develop valid statements that can be made about the preparedness of twelfth-grade students in NAEP reports.

Discussion of Assessment-to-Assessment Alignment

The study described in this report followed the alignment methodology documented in the paper by Dr. Norman Webb titled "Design of Content Alignment Studies in Mathematics and Reading for 12th Grade NAEP Preparedness Research Studies." The full document is included in Appendix A.

The Webb alignment methodology was originally designed to study the alignment between the standards on which a test is based and the test itself. That is, the original purpose of the Webb alignment methodology and software was not to compare two assessments to one another. At the Governing Board's request, Dr. Webb adapted the methodology to be used to study the alignment of two tests.

In an alignment study looking at how strongly a set of standards and a test are aligned, the Webb methodology requires that expert panelists make judgments about the cognitive complexity of the individual standards and of the test items, and it requires that the panelists determine whether each test item may be coded to (aligned with) a standard. Once these judgments are made, the data are analyzed and organized around four primary criteria: Categorical Concurrence, Depth-of-Knowledge Consistency, Range-of-Knowledge Correspondence, and Balance of Representation, all of which are discussed in more depth later in this report. For each criterion, statistical parameters are established that are used to indicate the relative strength with which the test alignment meets the criterion.

Adapting the methodology to study the alignment of two tests involves more steps in the process. To study the alignment of hypothetical Test A and Test B with one another, expert panelists must determine the cognitive complexity of the standards on which both tests are based as well as the complexity of all test items included in the study. Then, in four sub-studies, the panelists must determine 1) whether each item of Test A may be coded to a standard for Test A, 2) whether each item of Test A may be coded to a standard for Test B, 3) whether each item of Test B may be coded to a standard for Test A, and 4) whether each item of Test B may be coded to a standard for Test B.

Once the judgments are made for each of the four sub-studies, the degree of alignment for each sub-study is analyzed, using the same four alignment criteria that are used for single-test alignment studies. Finally, the statistical results of the four sub-studies are considered as a whole, and statements and comparisons are identified that illustrate the degree to which the content of the two tests is aligned.

Thus, the alignment methodology used for this study was designed to address similarities and differences between the content and skills measured by the NAEP and WorkKeys reading assessments, as well as the cognitive complexity of these assessments.

Methodology

Study Design

The Webb alignment methodology used for this study specifies that, prior to assembling the content experts for the alignment study, an independent content expert should conduct an analysis of the test blueprints. Accordingly, an expert in reading first analyzed the NAEP and WorkKeys test blueprints to identify similarities and differences in the respective tests' specifications. This analysis found that, while the NAEP assessment covers a broader range of cognitive targets than the WorkKeys assessment does, when the cognitive targets specifically associated with literary reading passages (fiction, literary nonfiction, poetry, and exposition) are removed from consideration, the remaining cognitive targets — those related to reading in general and to informative text specifically — are largely covered by both assessments. The full report on the blueprint analysis is included in Appendix B. Table 1 shows a comparison of the critical features of the frameworks and specifications for the NAEP Grade 12 *Reading* assessment and the WorkKeys *Reading for Information* assessment.

Table 1: Comparison of the critical features of the NAEP Grade 12* Reading *test and the WorkKeys* Reading for Information *test, excerpted from blueprint analysis report

	NAEP GRADE 12 READING ASSESSMENT	WORKKEYS *READING FOR INFORMATION* TEST
Types of Reading Passages	***Literary texts (30%)*** • *20% Fiction: e.g.,* adventure, historical fiction, realistic fiction, folktales/legends/myths/fantasy, satire, parody, allegory, monologue; intact passages or excerpts • *5% Literary nonfiction: e.g.,* personal essay, autobiographical/biographical, sketches, speech, character sketches, memoir, classical essay; intact passages or excerpts • *5% Poetry: e.g.,* narrative poem, free verse, lyrical poem, humorous poem, ode, song, epic, sonnet, elegy; intact poems or excerpts ***Informational texts (70%)*** • *30% Exposition: e.g.,* essay, literary analysis; intact passages or excerpts • *30% Argumentation or persuasive text: e.g.,* informational trade book, journal, speech, persuasive essay, letter to the editor, argumentative essay, editorial, historical account, position paper (brochure, campaign literature, advertisement, etc.) • *10% Procedural texts and documents: e.g.,* graphics and other information embedded in text, as well as stand-alone documents like applications, manuals, product support materials, and contracts ***Mixed texts*** ***Paired texts***	***Literary texts (0%)*** ***Informational texts (100%)*** • *100% Procedural texts and documents: e.g.,* contracts, policies, instructions, legal documents, information memos, letters, signs, bulletins, regulations, notices, directions. These documents may have some elements of argumentation or persuasion, particularly at Levels 6 and 7.

	NAEP GRADE 12 READING ASSESSMENT	WORKKEYS *READING FOR INFORMATION* TEST
Character-istics of Reading Passages	• Well organized, sufficient elaboration of new concepts, use of graphic features (italics, bold print, signal words and phrases) • High quality • Authentic • Coherent • Grade appropriate • Drawn from a variety of contexts • Engaging • Reflecting our literary heritage, including works from varied historical periods • Reviewed for potential bias and sensitivity issues	• Authentic texts from the work world • Coherence varies • Passages range from Level 3 to Level 7, with approximate reading levels of stimulus text from 6th grade to post high school • Drawn from 6 World of Work Career Clusters • Relevant • Reviewed for potential bias and sensitivity issues
Length of Reading Passages	Approximately 500–1,500 words: Passages of varying lengths are used in order to gain the most valid information about students' reading skills by reflecting the types of materials they encounter in and out of school. In addition, passages must be long enough to yield at least 10 associated test items.	Approximately 70–500 words; longer and more complex at higher levels
Reading Difficulty	• Primarily selected by expert judgment according to criteria described in the test specifications • Grade 12-appropriate reading level that includes a range of sentence and vocabulary complexity; at least two research-based readability formulas and passage mapping are used as selection guides	• Selected by expert judgment • Approximations: Level 3: 6th grade Level 4: 8th grade Level 5: 10th grade Level 6: 12th grade Level 7: post high school
Cognitive Targets	• Three levels of cognitive targets are addressed by NAEP items: – Locate/Recall – Integrate/Interpret – Critique/Evaluate • These cognitive targets are applied to the following categories of text: – Literary text – Informational text – Both literary and informational text • See Appendix B for full list of cognitive targets	• Five strands of reading skills are addressed by WorkKeys items: – Choosing main ideas or details – Understanding word meanings – Applying instructions – Applying information – Applying reasoning • Examinees are asked to apply these strands of skills across a variety of document types and at varying levels of complexity • See Appendix B for full list of skills

	NAEP GRADE 12 READING ASSESSMENT	WORKKEYS *READING FOR INFORMATION* TEST
Vocabulary-Related Tasks	• Identifying and understanding meanings of words within context • Application of understanding of word meanings to passage comprehension • Understanding how words convey concepts, ideas, actions, or feelings known by readers • Understanding how words are linked to the central idea and are necessary for understanding the context • Excluded words: words related to specific content domains; words that name or label the main idea; words in everyday speaking vocabulary; words explicitly defined in appositives, parentheses, etc.; jargon and technical terms	• Select definition from options provided • Identify definition by context • Identify definition by context of words with multiple meanings • Understand jargon and/or technical terms • Understand uncommon jargon and technical terms from context • Identify the meaning of an acronym that is defined in the passage • Figure out the meaning of an acronym that is not directly defined
Number of Items	• 10–12 items per passage, two vocabulary • Passage and items constitute a "block" • 131 total reading items in the NAEP pool; no single student completes all 131 items • Assessment of an individual contains two blocks: 20–24 items total • 20%–30% of items are intertextual	• 30 operational items (1, 2, or 3 items per passage) allocated at 6 items per level (Levels 3–7) and distributed across 6 Career Clusters and including all stimuli types: • 1–3 contract • 5–9 policy • 11–15 instructions • 1–3 legal documents • 4–8 information • Additional 3 items for pretest only (unscored) • An assessment for an individual consists of all items
Item Types	***3–6 multiple choice*** • 4 answer options: 1 correct, 3 incorrect • Assumed time to complete: 1 minute ***5–8 short constructed response*** • 1- or 2-sentence response • Assumed time to complete: 2 to 3 minutes ***1 extended constructed response*** • 1- or 2-paragraph response • Assumed time to complete: 5 minutes	***Multiple choice*** • 5 answer options: 1 correct, 4 incorrect

	NAEP GRADE 12 READING ASSESSMENT	WORKKEYS *READING FOR INFORMATION* TEST
Time Per Item Type	• 40% multiple choice (1 minute each) • 45% short constructed response (2–3 minutes) • 15% extended constructed response (5 minutes each) • 60% of time on constructed responses	• 100% multiple choice (5 answer choices)
Assessment Time	• 2 blocks at 25 minutes each • 50 minutes total	• 45 minutes paper and pencil • 55 minutes computer
When Given	Every 4 years, late January through early March	On demand
Testing Population	• Representative national sample of 8,000–10,000 12th-grade students per subject across the nation (about 200–300 schools) • The samples of students are designed to be representative of the nation and are drawn from different regions of the country and participating states • ELL students participate unless they have had less than 3 school years of instruction in English	• High school students, job applicants, current employees, people seeking certification or other documentation of their skill levels. Approximately 750,000 WorkKeys *Reading for Information* tests were administered in fiscal year 2009.

	NAEP GRADE 12 READING ASSESSMENT	WORKKEYS *READING FOR INFORMATION* TEST
Accommo-dations	• **Disallow** having passages or items read aloud • Allow 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 • For a complete list of NAEP reading accommodations see: http://nces.ed.gov/nationsreportcard/about/inclusion.asp#accom_table	• Word-for-word foreign-language dictionary • Approved translations • Extended time • Large print • Audio recording • Reader/signer script (exact English only) • Braille • Assistance in recording responses • Computer-based accommodations including special workstation configurations, magnification, and special mouse, but not screen readers
Item Scoring	• 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 • Students must support statements with information from the reading passage. • Responses are coded to distinguish between incorrect and blank responses. • Responses are scored on the basis of their content, not on the quality of writing.	• Multiple choice: – Incorrect 0 – Correct 1 • No penalty for guessing

	GRADE 12 NAEP READING ASSESSMENT	WORKKEYS *READING FOR INFORMATION* TEST
Test Scores	**Scaled scores:** Range of 0 – 500; average scores for groups **Achievement levels:** The numeric scale score range is divided into the following three achievement levels: • **Basic** — Partial mastery of prerequisite skills and knowledge necessary for proficient work • **Proficient** — 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, specifically: – find evidence in support of an argument – integrate information from a variety of sources – determine unstated assumptions – analyze point of view – judge the logic, coherence, or credibility of an argument – have sizable meaning vocabularies, knowledge of words beyond most common meanings, flexibility with word meaning to fit the different contexts and understand passage meaning • **Advanced** — Superior performance Test scores and achievement levels are used to report on the performance of groups of 12th-graders regionally, by state, and across the country.	Test scores are criterion referenced. **Level Scores:** Scores range from Level 3 to 7; a score of Below 3 also may be given. Level 3 is the lowest level generally useful in a job; individuals possessing reading skills below this level are generally not qualified for jobs that require even the most basic reading, and employers are typically not willing to train individuals with reading skills below this level. **Scale Scores:** Smaller units within each level score; these can be used to show increments of change over time. Test scores are provided to individuals. Individuals and employers can use test scores to compare individuals' skills to the skill levels required for particular jobs. Employers and educators can use test scores to determine skill gaps and target training to these gaps.

The results of the blueprint analysis informed the preparations for the full alignment study in several ways. Primarily, the blueprint analysis outlined general similarities and differences between the two assessments. This helped the contractors and facilitators to better prepare training and introductory materials for the panel participants. In addition, the blueprint analysis helped to inform decisions that were made about how to represent the two tests' standards for the study. A more thorough discussion of this process is included in the section of this report titled "Standards/Representation of the Domains."

The full study was planned for January 25-29, 2010, at the national headquarters of ACT, Inc., in Iowa City, IA. Per the Webb methodology, two concurrent, replicate panels would review the content representation and test items for both assessments and determine the extent to which the assessments measure similar content. Having two replicate panels conduct the alignment study concurrently would allow for a real-time check of the reliability of results. Comparable results from the two panels would indicate that confidence in the results is warranted.

The alignment methodology, described in greater detail by Dr. Webb in Appendix A, includes the following steps:

- Training two concurrent panels of content experts to conduct the analysis
- Assigning Webb's depth-of-knowledge (DOK) levels to test framework standards and objectives
- Assigning DOK levels and test framework objectives to test items
  - Map the NAEP items to the NAEP framework objectives
  - Map the WorkKeys items to the NAEP framework objectives
  - Map the NAEP items to the WorkKeys objectives
  - Map the WorkKeys items to the WorkKeys objectives
- Analyzing and reporting the results using four alignment criteria:
  - Categorical Concurrence
  - Depth-of-Knowledge Consistency
  - Range-of-Knowledge Correspondence
  - Balance of Representation

The Web Alignment Tool (WAT) was used to collect the data from panelists and for conducting the analyses. This is a Web-based software application designed by Dr. Webb to be used with his alignment methodology. All of the content standards for the two assessments are entered into the WAT. Then panelists enter DOK levels for standards and test items, as well as the test standards to which they believe the test items align. The WAT is programmed to perform alignment analyses on these data.

Pilot Study Lessons Learned

Prior to conducting the full alignment study, ACT conducted a smaller scale pilot study for the purpose of testing the alignment methodology and software so that the procedures could be fine-tuned in preparation for the full study to be held January 2010. The pilot study was held November 16 and 17, 2009, on the ACT campus in Iowa City, IA.

There were five participants in the pilot study: one facilitator and four panelists. The facilitator was one of the two facilitators selected for the full study in January, while the panelists were reading experts from Iowa who were not part of the full study in January. In addition, two ACT staff members and a representative from NAGB were present for the pilot.

The pilot used the same methodology and followed the same basic procedures planned for the full study in January, using a subset of the test items rather than the full item pools used in the full study. The subset of test items selected for use in the pilot was designed to be representative of the entire pool for the full study.

At the beginning of the two-day meeting, the participants received background information on the NAEP program, the WorkKeys system, and the NAGB preparedness research project of which this study is a part. Panelists were trained in the use of Depth of Knowledge (DOK) levels to indicate the complexity of the test framework objectives and test items. Panelists were also trained in the use of the Web Alignment Tool (WAT) software.

The pilot participants followed the same procedures intended for the full study, including training, group practice, independent analysis, group discussion and adjudication, and completing evaluation surveys about the procedures and the alignment. Panelists used the WAT software to record their independent judgments about the test framework objectives and test items. The data collected in the WAT software tool were analyzed solely to ensure that ACT understood how the analysis features of the WAT work; they were not analyzed for the purpose of evaluating the alignment of the two assessments because the pilot was only an abbreviated version of the full study.

The feedback received from the pilot participants via discussion and written evaluation forms was used to inform the preparations for the full study in January. Overall, the pilot confirmed that the methodology is solid and works as intended.

The primary lessons learned from the pilot and applied to the full study included these:

Background information — The participants desired additional background information on the context of the alignment study and the potential uses of the results. We determined that we should provide additional information about the NAEP and NAGB, the WorkKeys system, the research program of which this study is a part, and what steps the NAGB may take once the research program is concluded.

Technology — We gained experience in helping to ensure each participant's computer workstation worked smoothly, including general troubleshooting and creating a bookmark on each workstation for the WAT URL.

Training materials — We added a WorkKeys-specific example to the DOK training packet because none of the other examples in the packet were similar to the WorkKeys items but many were similar to the NAEP items.

Test framework representation — We determined, through discussion with the pilot panelists and, later, consultation with NAGB and ACT WorkKeys staff, that we should add descriptive text to the test framework representations of both tests at the "standard" level (top level of the outline) in order to clarify the standards for the panelists. For the pilot, there were only labels at this level of the framework representations (e.g., "Level 3" for WorkKeys; "Locate/Recall" for NAEP), and this was neither the best representation of the intention of the individual assessments' framework nor as clear as possible for the panelists.

Alignment study materials — The sheer volume of items made it challenging for the panelists to navigate the materials as they worked, so we determined that we should consolidate the pages associated with each NAEP item if possible. To do this, we removed the page that stated the correct answer for each multiple-choice item and instead wrote it in on the page with the item. In

addition, we ensured that the items were all numbered sequentially and that there were noticeable dividers between items, all to improve navigation among items.

Discussion and adjudication — The panelists felt strongly that full-group discussion was very important, particularly early in the process, as a means of standardizing training and helping participants to clarify their thinking about the process, the standards, and the DOK levels.

Questions of interpretation — The pilot study allowed us to predict that the following questions would receive a fair amount of attention from the participants in the full study:

1) Should DOK levels be influenced by grade level or individual capabilities, or is DOK strictly a criterion that is independent of such consideration?
2) How should standards or objectives that appear to incorporate elements of more than one DOK level be handled?
3) How should NAEP literary items be coded to the WorkKeys standards, which do not include literary writing?

For these questions, we concluded that the full study participants should determine how their groups would handle these issues.

Panel and Facilitator Qualifications and Criteria for Selection

NAGB required this alignment to be, and to have the appearance of being, independent of the tests under scrutiny to the maximum extent feasible. Toward that end, the alignment was to be conducted by a panel of experts, the majority of whom were not directly associated with either the WorkKeys or National Assessment of Educational Progress programs. The study was conducted according to a methodology developed independently for NAGB by Dr. Norman Webb and facilitated by independent consultants associated with Dr. Webb. However, the project was carried out under a contract with ACT, the developer and owner of the WorkKeys assessments, and this report was prepared by ACT staff. In addition, a list of potential panelists was provided by NAGB.

ACT recruited facilitators and panelists to participate in this study. The alignment methodology called for six to eight panelists in each of the two replicate panels. The panels were to be equivalent in terms of area of content expertise, level of content expertise (secondary/ postsecondary), and demographic attributes. Racial, ethnic, and geographic diversity was also recommended.

Panelists were recruited from universities, professional reading organizations, and professional networks. Special efforts were made to recruit individuals from typically underrepresented groups by contacting and requesting participation or referrals from the leadership of organizations that emphasize a diverse membership. Due to scheduling conflicts, the study included only one panelist who was a member of a minority racial or ethnic group. However, several panelists specialize in research and/or work focusing on literacy and diverse populations. The facilitators were recommended by Dr. Webb based on their extensive experience in working with him on many other alignment studies.

ACT obtained commitment from a total of 16 panelists and two facilitators for the study. After attrition and last-minute travel cancellations due to weather, the two panels included six and seven experts respectively, plus a facilitator for each panel.

Panel assignments were made to ensure that the two groups were roughly balanced in terms of gender, geography, background, and experience. (See Appendix T for brief biographies of project participants and staff.)

Standards/Representation of the Domains

The alignment methodology required that the test content for the two tests being studied be represented in a manner compatible with the Web Alignment Tool (WAT), the software tool designed by Dr. Webb for use with alignment studies. The alignment methodology refers to such a representation as the test *standards*. The WAT requires that standards being used for an alignment study be organized in an outline structure, with *standards* as primary headings and *objectives* beneath the standards in the outline. Although the documents from which the content representation was derived for the two assessments used in the study do not necessarily refer to them as "standards," this term will be used in this report for the purpose of simplicity. The documents that served as the standards for the study are in Appendix E. The text in the remainder of this section describes how the standards for the two tests were adapted from their respective content representation.

The version of the NAEP standards used for this study were approved by NAGB and based on Exhibit 8 from the *Reading Framework for the 2009 National Assessment of Educational Progress* (September, 2008; p. 39). This exhibit from the *Reading Framework* presents the cognitive target matrix for the test. The matrix includes three levels of cognitive targets: Locate/Recall, Integrate/Interpret, and Critique/Evaluate. For each cognitive target level, the exhibit lists knowledge and skills for each of the following categories: Both Literary and Informational Text, Specific to Literary Text, and Specific to Informational Text. In order to fit within the constraints of the WAT, Exhibit 8 from the NAEP Framework document was translated into an outline format.

Following is an excerpt from the NAEP Grade 12 *Reading* assessment standards used in this study. Throughout this report, the text describing the cognitive level is at the top level of the outline excerpted in Table 2 (e.g., 1) and is referred to as the generic *standard*. The text at the second level of the outline (e.g., 1.1) tells the type of text to which each objective is applied. The text describing specific cognitive targets is at the third, lettered level of the outline (e.g., 1.1.a) and is referred to as the *objective*.

Table 2: Excerpt from NAEP Grade 12* Reading *standards

Level	Description
1	Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension
1.1	Locate or recall textually explicit information and make simple inferences within and across *both literary and informational texts*
1.1.a	Locate or recall specific information such as definitions, facts, and supporting details in text or graphics
1.2	Locate or recall textually explicit information and make simple inferences within and across *literary texts*
1.2.a	Locate or recall character traits
1.2.b	Locate or recall sequence of events or actions
1.2.c	Locate or recall setting
1.2.d	Locate or recall figurative language
1.2.e	Locate or recall organizing structures of literary texts, such as verse or stanza in poetry or description, chronology, comparison, etc. in literary non-fiction

The WorkKeys *Reading for Information* framework is organized by increasing cognitive complexity from one skill level to the next and covers a range of skills typically required in the workplace. There are five skill levels of increasing cognitive complexity, ranging from 3 to 7. Characteristics of the reading passages vary in complexity according to the skill level, and the skills required by test items parallel the complexity of the reading passages with which they are associated. The WorkKeys *Reading for Information* framework was structured for use in this study so that the overall skill level descriptors were the standards. The characteristics of items at each skill level were used as the objectives.

The version of the WorkKeys *Reading for Information* test standards used for this study was adapted from the WorkKeys Skill Definitions source documents. The adaptation was made so that the format of the WorkKeys standards would conform to the requirements of the WAT. Internal subject matter experts for the *Reading for Information* assessment and the WorkKeys program were consulted to ensure that the adapted standards were consistent with the original skill definitions.

Readers will note that the WorkKeys standards begin at Level 3, rather than 0 or 1. ACT recognizes that there are levels of reading skill below what is represented by WorkKeys Level 3. However, ACT's research during the development of the WorkKeys system showed that skills at these levels are lower than the skills that employers are typically willing to accept. Thus, the WorkKeys scale begins at Level 3 both as acknowledgement that lower-level skills do exist and as recognition of the lowest commonly accepted skill levels for the workplace.

Following is an excerpt from the WorkKeys *Reading for Information* assessment standards used in this study. Throughout this report, the text at the top level of the outline excerpted in Table 3 (e.g., 3) is referred to as the generic *standard*. The text at the second level of the outline (e.g., 3.1) is referred to as the *objective*.

Table 3: Excerpt from WorkKeys* Reading for Information *standards

Level	Description
3	Clearly stated, simple information; elementary vocabulary; no conditional statements
3.1	Apply instructions to a situation that is the same as the one in the reading materials
3.2	Choose the correct meaning of a word that is clearly defined in the reading
3.3	Choose the correct meaning of common, everyday workplace words
3.4	Choose when to perform each step in a short series of steps
3.5	Identify main ideas and clearly stated details
4	Clearly stated, detailed information; common workplace vocabulary; may have conditional statements
4.1	Apply instructions with several steps to a situation that is the same as the situation in the reading materials
4.2	Choose what to do when changing conditions call for a different action (follow directions that include "if-then" statements)

Assessments

The NAEP Grade 12 *Reading* Assessment: The NAEP items to be used for this study were organized into blocks. The NAEP program uses a matrix sampling procedure to construct forms that can be administered feasibly under classroom conditions. This process means that no single test form is a representative sample of the breadth of items that could appear on a form. Therefore it was necessary to examine the entire pool of 131 items for the 2009 assessment provided by NAGB for this study (items from the NAEP vocabulary blocks are not included in this analysis because they are not included in the NAEP Grade 12 reporting scale).

The items for the 2009 NAEP assessment used for this study included both multiple-choice (76 items) and constructed-response items (55 items), and they were associated with 15 reading passages. Thirty percent of the texts were literary — including fiction, poetry, and literary nonfiction — and 70 percent were informational — including informational expository, argumentative/persuasive, and procedural texts. Three of the 15 NAEP reading passages used for this study had an explicit workplace context, one was oriented toward the consumer, and the remaining passages did not have an explicit workplace context. The scoring for the NAEP test uses item-response theory, and scoring rubrics are used for the constructed-response items that have point totals from 1 to 4. However, for the purposes of this study, all items were weighted equally.

To enter the NAEP items into the WAT, it was necessary to number them sequentially. Each numbered item in the WAT was labeled with the block and sequence information so that it could be tied back to the original NAEP item. Sequential numbering also helped ensure that the panelists would enter their data for the item they intended.

The WorkKeys *Reading for Information* Assessment: All WorkKeys assessments are constructed according to a blueprint that specifies the type and number of items they will contain. This blueprint is designed, in part, to ensure that all test forms for a particular content area have parallel content and equivalent difficulty.

ACT provided two intact WorkKeys *Reading for Information* forms consisting of 30 operational items each, with 60 unique items in all. Items on the second form were renumbered so that they would be in sequence following the first form. The item pool used for this study included 28 documents with which one to three test items were associated. All documents used in WorkKeys *Reading for Information* assessments have a workplace context, and they include policy, instructions, legal documents, and informational texts. All WorkKeys reading test items are multiple choice and worth 1 point. WorkKeys test forms are parallel and equated using items from a pool of hundreds of operational items.

Materials and Preparation

Prior to the study, participants received and were required to review a general description of the study, the NAEP framework, the WorkKeys *Reading for Information* technical manual, and an agenda for the week. A lead facilitator was selected to conduct the training for both panels, ensuring that they would be operating from the same foundation.

In addition, all participants and staff for the study were bound by confidentiality and nondisclosure agreements that required them to use all confidential, proprietary materials for the purposes of the study only. During the on-site panel meetings, participants were required to adhere to strict security policies and procedures that included keeping personal items such as purses, bags, and cell phones away from their work spaces. NAEP and WorkKeys test materials were guarded or kept in locked locations when not in use, and all confidential materials were returned to ACT staff for secure storage and subsequent secure destruction at the end of the study.

Materials prepared for the on-site meeting included the following:

- Test items: Each participant received a binder with all NAEP and WorkKeys test items and scoring rubrics (for NAEP constructed-response items). The binders were securely stored in a locked location whenever the panelists were not using them.
- Test standards: Each participant received printed copies of the standards that had been entered into the WAT.
- Evaluation forms: Dr. Webb's alignment methodology specifies that evaluation surveys be completed by panelists after many steps of the alignment process. NAGB requested that ACT use the same forms as used by another contractor for a related study, and this was done. Evaluation forms are included in Appendix F. Panelists' responses to the evaluation forms are found in Appendix G.
- Training packet: Training materials related to Dr. Webb's depth-of-knowledge (DOK) levels were adapted from Dr. Webb's alignment materials, with his assistance. This training packet is found in Appendix D.
- Meeting facilities: Arrangements for the panel meetings were made in the ACT conference center. One large room with a divider was used to allow both large- and small-group discussion. Each panelist was provided with a desktop computer with Internet service for access to the WAT.

Additionally, prior to the beginning of the on-site meetings, ACT personnel registered the two concurrent panels in the WAT, uploaded the standards and assessments, and created four studies

within the WAT: NAEP assessment to NAEP standards, WorkKeys assessment to NAEP standards, NAEP assessment to WorkKeys standards, and WorkKeys assessment to WorkKeys standards.

Procedure

On the first day of the week-long panel meetings, after introductions and administrative details were covered, a representative from NAGB presented a context for this alignment and an overview of the NAEP. A representative from ACT provided an overview of the WorkKeys system, with special focus on the *Reading for Information* assessment.

The lead facilitator then presented the training to the two panels combined. This was done to ensure uniformity of training. The trainer described the alignment methodology in general and described in detail the process specific to the reading content area. The trainer then guided the panelists through a general overview of the Depth-of-Knowledge (DOK) levels, followed by specific training on DOK as applied to reading. Last, the panelists independently practiced labeling sample items with DOK levels, then discussed their judgments as a large group. This allowed the full group of panelists to achieve a common understanding of the DOK levels and how to accurately and consistently apply them to the reading content area. Finally, panelists evaluated the quality of the presentations and training. All evaluation results are in Appendix G.

Once the training was complete, the facilitators received their WAT registration logins, group number assignments, and passwords. Panelists registered with their respective groups. Each participant received the NAEP standards, WorkKeys standards, and a binder containing the items from each assessment. The binders were securely locked when participants were not using them. After completing each study and at the end of each day, panelists completed an evaluation form specific to the activity completed.

Each panel performed the following tasks, as specified by the Webb alignment methodology (for a detailed description of all steps, see Appendix A):

Sub-Study 1: NAEP Grade 12 Reading *items to NAEP Grade 12* Reading *standards*

- Assign DOK levels to each objective in the NAEP standards
- Adjudicate within each panel to achieve consensus on DOK levels
- Facilitators identify and adjudicate differences between the two groups to achieve inter-panel consensus on DOK levels
- Assign DOK levels to NAEP items
- Map NAEP items to the NAEP standards
- Adjudicate mapping within each panel
- Complete evaluation of just-completed work

Sub-Study 2: WorkKeys Reading for Information *items to NAEP Grade 12* Reading *standards*

- Assign DOK levels to WorkKeys items
- Map WorkKeys items to NAEP standards
- Adjudicate mapping within each panel
- Complete evaluation of just-completed work

Sub-Study 3: NAEP Grade 12 Reading *items to WorkKeys* Reading for Information *standards*

Assign DOK levels to each objective in the WorkKeys standards
Adjudicate within each panel to achieve consensus on DOK levels
Facilitators identify and adjudicate differences between the two groups to achieve inter-panel consensus on DOK levels
Map NAEP items to the WorkKeys standards
Adjudicate mapping within each panel
Complete evaluation of just-completed work

Sub-Study 4: WorkKeys Reading for Information *items to WorkKeys* Reading for Information *standards*

Map WorkKeys items to WorkKeys standards
Adjudicate mapping within each panel
Complete evaluation of just-completed work

Debriefing

Discussion
Written evaluation of overall alignment process and results; recommendations regarding the alignment and appropriate uses of results

The two facilitators communicated throughout each day's work and also in the evenings to identify areas their respective panels should discuss further and to plan any necessary adjustments to the procedures.

Decision Rules and Adjudication

Due to the unique characteristics of the NAEP and WorkKeys assessments, the demands of a test-to-test alignment, the interaction of the panelists, and time constraints, there were some variances from the prescribed alignment methodology. In addition, the panels found it necessary to establish decision rules in some situations where differences between the two assessments would have led to an inconclusive variety of judgments among the panelists without the consensus and guidance of the decision rules. Decision rules helped panelists avoid ambiguous situations that may have been confusing and inefficient. The variances, decision rules, and rationales for each follow:

1) Decision rule, coding NAEP items to NAEP standards: For constructed-response items, use the rubric for the answer with the highest point value when determining how to code items.

Rationale: The rubric for the answer with the highest point value reflects the full intended content and cognitive demand for the item. This approach is analogous to considering the thought that is required to respond correctly to a multiple-choice item.

2) Agenda adjustment, Wednesday, January 27, 2010: Adjudicate discrepant NAEP item-to-NAEP standards coding after completion of coding WorkKeys items to NAEP standards.

Rationale: The agenda called for completing the NAEP item-to-NAEP standards coding on Tuesday, including adjudication. However, due to the large number of NAEP items, the

adjudication was not completed Tuesday. In the interest of staying as close to the agenda as possible, the facilitators agreed to begin Wednesday according to the original agenda. All felt that the WorkKeys items-to-NAEP standards coding would not take the full amount of time allotted on Wednesday, and also that the WorkKeys standards DOK coding would not take the full time allotted. Therefore, it seemed reasonable to expect that there would be some extra time available Wednesday afternoon for adjudicating the NAEP items-to-NAEP standards coding.

3) Adjudication procedure: The groups would identify the items for which the panelists' judgments are most discrepant and begin the adjudication with those. They would then work "backward" toward the items that were least discrepant.

Rationale: The facilitators established this approach in response to the fact that time was very limited. In the event that there was not enough time to adjudicate all of the discrepant judgments, this approach would ensure that at least the most discrepant judgments were adjudicated by the group, knowing that the Webb methodology allows for adjudication by the facilitators in cases where adjudication with the panelists has not been completed.

4) Decision rule, coding WorkKeys items to NAEP standards: For some WorkKeys items that involve sequencing of events or actions, panelists will code the items to NAEP standard 1.2.b, "Locate or recall textually explicit information and make simple inferences within and across *literary texts* — Locate or recall sequence of events or actions." Along with coding to this standard, panelists will include a note in the WAT that explains that the WorkKeys item does not refer to a literary text.

Rationale: NAEP standards refer to sequence only in the standards that are specific to literary texts. Many WorkKeys stimulus passages involve sequencing of some sort, but none of the passages are literary. If this decision rule had not been made, there would be a large number of WorkKeys items that would be marked "uncodable." Therefore, because the panelists felt that the basic task of locating or recalling sequence of events or actions in text is the same in both literary and informational texts, the decision was made to code applicable WorkKeys items to NAEP standard 1.2.b in order to avoid creating the impression of a larger lack of alignment than there was in reality. Table 12 indicates that 11 – 12% of the WorkKeys items were coded to standard 1.2.b.

5) Decision rule, coding NAEP items to WorkKeys standards: When coding NAEP items to WorkKeys standards, the difficulty of the NAEP stimulus passage will be taken into account.

Rationale: This approach mirrors how the WorkKeys test is constructed. The WorkKeys stimulus passages and associated items are intended to be at the same skill level.

6) Note template: Panelists were given the following "Note" template to use when coding NAEP items for which there is no corresponding WorkKeys standard: "This item assesses _____, which is not addressed in the WorkKeys objectives."

Rationale: In a typical alignment using Webb's methodology and the WAT, it is uncommon for an item to be uncodable. However, the particular differences between the NAEP and WorkKeys standards are such that panelists and facilitators agreed in discussion that a significant number of NAEP items would not be codable to the WorkKeys objectives.

The NAEP standards are generally organized in a hierarchy of cognitive skills, with objectives within each standard organized according to type of reading passage: literary, informational, or both literary and informational. In contrast, the WorkKeys standards are organized in a hierarchy of increasingly complex cognitive skills, all of which are applied to informational texts.

In a typical alignment, when panelists code an item for which there is not a clear matching objective, they have the option to code the item to the standard at the head of a group of objectives (the "generic" standard). However, the panelists concluded that such coding practice may be misleading, indicating alignment that was not actually present. For example, there is no WorkKeys standards category that could accurately be applied to a NAEP item about a literary device such as theme.

The WAT requires that, if panelists code an item as "uncodable," they must type in a note explaining why it is uncodable.

Recognizing that A) it was likely that the panelists would code many NAEP items as uncodable to the WorkKeys standards, and B) that having to compose a new note each time would be very time consuming over the course of all of the NAEP items, the facilitators provided a template note for panelists to copy each time they determined a NAEP item was uncodable. The panelists were required to include variable text of what the particular NAEP item did assess, along with the standard text: "This item assesses _____, which is not addressed in the WorkKeys objectives."

7) Decision rules, coding NAEP items to WorkKeys standards — uncodable items: Both panels agreed to regard the following types of NAEP items as uncodable to WorkKeys standards:

- NAEP items related to literary devices such as theme
- NAEP items related to author's craft
- NAEP items that require the examinee to construct a response to explain, critique, or evaluate something

Only one of the two panels agreed to consider the following type of NAEP item as uncodable to WorkKeys standards:

- NAEP vocabulary items with a DOK level of 1 and associated with a literary stimulus passage

8) Decision rule, coding NAEP items to WorkKeys standards: If the WorkKeys objective coded to a given NAEP item is at a significantly lower skill level than the NAEP reading passage, the panelists will include a note in the WAT about the reason the objective was selected.

Rationale: The WorkKeys blueprint specifies that items should be at the same skill level as the associated reading passage, and the WorkKeys objectives reflect this specification. Therefore, panelists felt it was necessary to provide justification whenever there was significant skill-level discrepancy among a NAEP reading passage, an associated NAEP item, and a WorkKeys objective coded to the item.

9) Process change, coding WorkKeys items to WorkKeys standards: Both groups were brought together to discuss decision rules before coding WorkKeys to WorkKeys, rather than determining decision rules independently.

Rationale: When the independent panels began coding WorkKeys to WorkKeys, they each started making decision rules, and the facilitators realized that the rules being discussed had the potential to be quite different from one another. In order to help prevent systematic, large differences between the two groups' coding, the facilitators decided to have a large-group discussion to bring the two panels to common ground, thus avoiding some of the anticipated discrepancies

10) Decision rule, coding WorkKeys items to WorkKeys standards: When coding the WorkKeys items, begin by first determining the level of the associated text, and then evaluate the cognitive demand required to process the item in relation to the text.

Rationale: The WorkKeys blueprint specifies that items should be at the same skill level as the associated reading passage, and the WorkKeys objectives reflect this specification. Therefore, panelists needed to consider the skill levels of both the item and the associate passage when coding a WorkKeys item. This decision rule was established to provide a uniform process for the panelists to follow.

11) Decision rule, coding WorkKeys items to WorkKeys standards: When coding items associated with text that contains instructions, the objectives related to instructions will be used only when the item clearly requires the examinee to apply the instructions to a conditional or novel situation not exactly addressed in the text. Objectives related to details will be used for items that require examinees to locate or identify specific information, even if the information is within a set of instructions.

The application of this decision rule was not entirely straightforward, however. Both panels struggled with this issue and found it difficult to apply the decision rule consistently. Despite having the decision rule, both panels found it necessary to adjudicate the coding for some items that fell in this category.

Alignment

As described by Dr. Webb, "Alignment … generally attends to the agreement in content between state curriculum standards and state assessments. 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." (Webb, 1997, p. 3)

In the case of this particular alignment study, an additional dimension is examined. In addition to analyzing the degree of alignment between a set of standards and the assessment based on that set of standards, the degree of alignment between two different assessments is examined. This is accomplished by evaluating the degree to which the test items align to the standards on which they are based, as well as evaluating the degree to which they are aligned with the standards for the other test.

It is important to point out 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. Alignment is intimately related to test "validity," most closely with content validity and consequential validity (Messick, 1989 [*sic*], 1994; Moss, 1992). Whereas validity refers to the appropriateness of inferences made from information produced by an assessment (Cronbach, 1971), content alignment refers to the degree to which content coverage is the same between an assessment and other curriculum documents (Webb, 2009, p. 2).

Alignment Criteria Used for This Analysis

Norman Webb's alignment methodology uses four criteria to determine the degree of alignment between standards and assessments.

- Categorical Concurrence: When applied to the alignment between a test and the standards on which it is based, this criterion measures the extent to which the same categories of content appear in the standards and the test items. A given standard is considered to be fully assessed by a test if there are at least six assessment items targeting that standard. Thus, this criterion is sensitive both to the total number of items and to the total number of standards evaluated for a given test.

  When applied to the alignment between two assessments, Categorical Concurrence refers to the extent to which the same categories of content are measured by both assessments.

  For this study, if there are six or more items targeting a given standard, the WAT indicates "Yes," the Categorical Concurrence alignment criterion has been met for that standard; if there are five items, the WAT indicates that the alignment is "Weak"; and when four or fewer items target a given standard, the WAT indicates "No," the Categorical Concurrence criterion has not been met for that standard. (WAT Training Manual, p. 110)

- Depth-of-Knowledge Consistency: When applied to the alignment between standards and an assessment, this criterion measures the degree to which the knowledge elicited from examinees on the assessment is as cognitively complex as what is stated in the standards. The criterion is met if at least half of the objectives in a standard are targeted by items of the appropriate complexity.

  When applied to the alignment between two assessments, Depth-of-Knowledge Consistency indicates whether the same depth of content knowledge is elicited from examinees by both assessments.

  For this study, if at least 50% of the items targeting a standard are at or above the DOK level of the objective to which they align, the WAT indicates "Yes," the Depth-of-Knowledge Consistency criterion has been met for that standard; if 41% – 49% of the items targeting a standard are at or above the DOK level of the objective to which they align, the WAT indicates that the alignment is "Weak"; and if 0% – 40% of the items targeting the standard are at or above the DOK level of the objective to which they align, the WAT indicates "No," the Depth-of-Knowledge Consistency criterion is not met for that standard. (WAT Training Manual, p. 111)

- Range-of-Knowledge Correspondence: This criterion measures whether the span of knowledge expected of examinees on the basis of a standard corresponds to the span of knowledge that examinees need in order to respond correctly to the corresponding assessment items or activities. The criterion is met for a given standard if at least half of the objectives that fall under that standard are targeted by at least one test item. Therefore, this criterion is sensitive to the total number of items evaluated for a given test, as well as to the number of objectives listed for each standard. For instance, if there is a small number of items in the item pool being studied, this may cause range-of-knowledge consistency to be weak. Similarly, if there is a large number of objectives listed for a given standard, this may cause the range-of-knowledge consistency to be weak for that standard.

  When applied to the alignment between two assessments, this criterion refers to whether a comparable span of knowledge within topics and categories is targeted by both assessments.

  For this study, if at least one test item aligns to at least 50% of the objectives within a standard, the WAT indicates "Yes," the Range-of-Knowledge Correspondence criterion is met for that standard; if at least one test item is aligned to 41% – 49% of the standards within an objective, the WAT indicates that the alignment is "Weak"; and if at least one item aligns to 0% – 40% of the objectives within a standard, the WAT indicates "No," there is not alignment using the Range-of-Knowledge Correspondence criterion. (WAT Training Manual, p. 112)

- Balance of Representation: This criterion measures whether the degree to which an objective is emphasized by test items is the same degree to which the objective is

emphasized in the standards on which the test is based. It evaluates whether items aligned to a given standard are clustered on just a few objectives, or they are spread among all objectives within the standard. Webb further explains: "An index value of 1 signifies perfect balance and is obtained if the corresponding items related to a content category [or standard] are equally distributed among the course-level expectations [or objectives] for the category. Index values that approach 0 signify that a large proportion of the items only correspond to one or two of all of the subcategories with at least one assigned item."

When applied to the alignment between two assessments, this criterion indicates whether a similar emphasis is given to the content topics and subtopics on each assessment.

For this study, if an index value is calculated to be 0.7 or higher, the WAT indicates "Yes," the Balance of Representation alignment criterion has been met; if the index value is 0.61 to 0.69, the WAT indicates that the alignment is "Weak"; and if the index value is 0.60 or less, the WAT indicates "No," the Balance of Representation alignment criterion has not been met for that standard. (WAT Training Manual, pp. 112 – 113)

Depth-of-Knowledge Levels

The explanation of Depth-of-Knowledge (DOK) levels in this section is taken from the materials developed by Dr. Webb:

Four DOK levels were used to judge both reading and writing objectives and assessment tasks. The reading levels are based on Valencia and Wixson (2000, pp. 909–935).

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

- Support ideas by reference to verbatim or only slightly paraphrased details from the text
- Use a dictionary to find the meanings of words
- Recognize figurative language in a reading passage

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

However, items require closer understanding of text, possibly through the item's paraphrasing both of both the question and the answer. Some examples that represent, but do not constitute all of, Level 2 performance are:

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

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

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

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

- Analyze and synthesize information from multiple sources
- Examine and explain alternative perspectives across a variety of sources
- Describe and illustrate how common themes are found across texts from different cultures

(Depth-of-Knowledge Levels section taken from Webb, 2005, pp. 70-71)

Results

Rater Data

Table 4 shows analysis of rater agreement for all four sub-studies. In each cell, the first two values are related to rater agreement for coding DOK levels to test items. If the intraclass correlation value is greater than 0.7, the correlation is considered to be adequate, and where it is greater than 0.8, it is considered to be good. The pairwise agreement is also calculated for coding DOK levels to test items, in case very low variance between the items has caused the intraclass correlation to be falsely high. For this statistic, a value of 0.6 indicates reasonable agreement and a value of 0.7 or higher indicates good agreement. Values of less than 0.5 indicate poor agreement. (WAT Training Manual, p. 116)

The third and fourth values in each cell are related to rater agreement for assigning test objectives to items. As explained earlier, the test content for this study has been organized into an outline structure, with *standards* as primary headings and *objectives* beneath the standards in the outline. The statistics in this table show interrater agreement at both the objective (detail) and the standard (broader) level.

Table 4: Rater agreement statistics for all four sub-studies

Sub-Study	Panel 1	Panel 2
Sub-Study 1: NAEP to NAEP	*Intraclass Correlation:* 0.95 *Pairwise Comparison:* 0.69 *Objective Pairwise Comparison:* 0.55 *Standard Pairwise Comparison:* 0.80	*Intraclass Correlation:* 0.95 *Pairwise Comparison:* 0.75 *Objective Pairwise Comparison:* 0.63 *Standard Pairwise Comparison:* 0.83
Sub-Study 2: WorkKeys to NAEP	*Intraclass Correlation:* 0.92 *Pairwise Comparison:* 0.80 *Objective Pairwise Comparison:* 0.86 *Standard Pairwise Comparison:* 0.93	*Intraclass Correlation:* 0.96 *Pairwise Comparison:* 0.87 *Objective Pairwise Comparison:* 0.87 *Standard Pairwise Comparison:* 0.95
Sub-Study 3: NAEP to WorkKeys	*Intraclass Correlation:* 0.91 *Pairwise Comparison:* 0.64 *Objective Pairwise Comparison:* 0.75 *Standard Pairwise Comparison:* 0.81	*Intraclass Correlation:* 0.95 *Pairwise Comparison:* 0.75 *Objective Pairwise Comparison:* 0.83 *Standard Pairwise Comparison:* 0.84
Sub-Study 4: WorkKeys to WorkKeys	*Intraclass Correlation:* 0.92 *Pairwise Comparison:* 0.76 *Objective Pairwise Comparison:* 0.80 *Standard Pairwise Comparison:* 0.89	*Intraclass Correlation:* 0.93 *Pairwise Comparison:* 0.82 *Objective Pairwise Comparison:* 0.87 *Standard Pairwise Comparison:* 0.90

For further explanation of the rater agreement statistics and how they were calculated, refer to Appendix U, Explanation of Rater Agreement Statistics.

As shown, there is just one sub-study for which one panel's pairwise comparison value is not in the "reasonable" or "good" range. For Sub-Study 1, Panel 1's pairwise comparison for objectives is 0.5467, which is in the "weak" range. This is typically due to overlapping objectives in the standards and/or items that may be answered using different approaches (e.g., using algebra or geometry in math) (see Appendix A: Alignment Methodology, pp. 19 – 20). Several panelists included comments in their coding for this sub-study about finding partial matches to one or more objectives and then deciding on one objective. The lower pairwise comparison value may be explained by this.

As may be expected, the interrater agreement is higher at the standard (broader) level than at the objective (detail) level. In general, Panel 2 shows somewhat stronger interrater agreement than Panel 1.

Table 4 shows that both panels demonstrate a high degree of interrater agreement. Thus, it is reasonable to have confidence in the reliability of each panel's ratings.

DOK Levels of the Standards

The methodology required the panels to reach inter-panel consensus on the DOK levels for each objective within the two tests' standards. Table 5 shows the DOK data for the NAEP standards.

Table 5: DOK data for the NAEP Grade 12* Reading *standards

NAEP Standard	# of Objectives	# and % of Obj. at DOK 1	# and % of Obj. at DOK 2	# and % of Obj. at DOK 3	# and % of Obj. at DOK 4	Average DOK
1.1	1	1 (100%)	-	-	-	**1**
1.2	5	4 (80%)	1 (20%)	-	-	**1.20**
1.3	4	-	4 (100%)	-	-	**2**
1 overall	10	5 (50%)	5 (50%)	-	-	**1.50**
2.1	6	-	1 (17%)	5 (83%)	-	**2.83**
2.2	5	-	-	5 (100%)	-	**3**
2.3	5	-	3 (60%)	2 (40%)	-	**2.40**
2.4	1	-	1 (100%)	-	-	**2**
2 overall	17	-	5 (30%)	12 (70%)	-	**2.71**
3.1	3	-	-	2 (67%)	1 (33%)	**3.33**
3.2	3	-	-	3 (100%)	-	**3**
3.3	4	-	-	4 (100%)	-	**3**
3 overall	10	-	-	9 (90%)	1 (10%)	**3.10**
ALL*	**37**	**5 (14%)**	**10 (27%)**	**21 (57%)**	**1 (3%)**	**2.49**

** Does not equal 100% due to rounding*

Table 5 shows that the NAEP standards associated with "locating and recalling" have an average DOK level of 1.50; the standards associated with "integrating and interpreting" have an average DOK level of 2.71, and the standards associated with "critiquing and evaluating" have an average DOK level of 3.10. It also shows that there is at least one objective at each of the four DOK levels.

Table 6 shows the DOK data for the WorkKeys standards.

Table 6: DOK data for the WorkKeys* Reading for Information *standards

WorkKeys Standard	# of Objectives	# and % of Obj. at DOK 1	# and % of Obj. at DOK 2	# and % of Obj. at DOK 3	# and % of Obj. at DOK 4	Average DOK
3	5	4 (80%)	1 (20%)	-	-	**1.20**
4	4	1 (25%)	3 (75%)	-	-	**1.75**
5	6	1 (17%)	5 (83%)	-	-	**1.83**
6	7	-	4 (57%)	3 (43%)	-	**2.43**
7	3	-	2 (67%)	1 (33%)	-	**2.33**
ALL	**25**	**6 (24%)**	**15 (60%)**	**4 (16%)**		**1.92**

Table 6 shows DOK levels for WorkKeys objectives ranging from 1 to 3. It also shows a general trend of increasing DOK level as the skill level of the WorkKeys standards increases from Level 3 to Level 7.

The DOK values of the individual standards for the two assessments range from 1 to 3.33 for the NAEP assessment, and from 1.20 to 2.43 for the WorkKeys assessment. On average, the DOK levels of the NAEP standards are higher than those for the WorkKeys standards, with the average NAEP standard DOK being 2.49 and the average WorkKeys standard DOK level being 1.92.

A factor in the difference in the average DOK level of the test standards is pointed out in the blueprint analysis. The Grade 12 NAEP reading passages are grade twelve-appropriate. In contrast, the workplace-focused WorkKeys reading passages range from approximately sixth-grade reading level to postsecondary level. The two panels struggled with whether and how to consider grade level as they determined DOK levels for the two assessments. Decision Rules 5, 8, and 10 are related to this issue; the questions of difficulty and skill levels to which they refer include the concept of grade level.

DOK Levels of the Test Items

In contrast to the test standards, the study methodology did not require consensus for the DOK levels of the test items. Nevertheless, the two panels reached similar conclusions about the DOK levels, and tables showing the DOK levels assigned by each panelist for each item are found in the appendices. The tables show that Panel 1 members assigned DOK levels to the NAEP items such that the average DOK for all NAEP items considered together was 2.16, while Panel 2 members assigned DOK levels such that the average was 2.14. Taking the assigned DOK levels of both panels together, the average for all NAEP items was 2.15.

For the WorkKeys items, the two panels show slightly more difference, but their averages are still similar. Panel 1 assigned DOK levels to the WorkKeys items such that the average DOK for all WorkKeys items considered together was 1.62. Panel 2 members assigned DOK levels that averaged to 1.47. Taken together, the average DOK level of all WorkKeys items across both panels was 1.54.

The following table summarizes the average DOK levels of the item pools studied for the two assessments.

Table 7: Average DOK levels of test items

	NAEP Items	WorkKeys Items
Average Item DOK Level	2.15	1.54

All WorkKeys *Reading for Information* items are multiple choice, whereas the NAEP assessment includes multiple-choice, short constructed-response, and extended constructed-response items, with the constructed-response items making up 42% of the NAEP item pool. The panels interpreted Webb's definitions of the DOK levels to mean that multiple-choice items could not be

coded to DOK Level 4, so no multiple-choice items for either test were coded to DOK Level 4. The Table 8 shows the average DOK levels of each item type used on the two tests.

Table 8: DOK levels by item type

Grade 12 NAEP *Reading* Test*						
Item Type	**# at DOK Level 1**	**# at DOK Level 2**	**# at DOK Level 3**	**# at DOK Level 4**	**Average DOK**	**Total # of Items**
Multiple choice	26	44	6	0	1.74	76
Short constructed response	3	9	30	0	2.64	42
Extended constructed response	0	1	12	0	2.92	13
WorkKeys *Reading for Information* Test*						
Item Type	**# at DOK Level 1**	**# at DOK Level 2**	**# at DOK Level 3**	**# at DOK Level 4**	**Average DOK***	**Total # of Items**
Multiple choice	30	30	0	0	1.54	60

** Rounding used for this table causes a slight discrepancy with some values used in the preceding report text. See tables in Appendices H and I for raw data and unrounded values.*

DOK Levels of Standards and Items Compared

When comparing the average DOK levels of the test standards with those of the test items, a similar pattern may be found: The average DOK level of the test standards for both assessments studied is higher than that of the average DOK level of the test items.

Specifically, for the NAEP assessment, the average DOK level for the standards was 2.49, and the average DOK level for the items was 2.15. The difference between average NAEP standard DOK level and average NAEP item DOK level was 0.34.

For the WorkKeys assessment, the average DOK level for the standards was 1.92, and the average DOK level for the items was 1.54. The difference between the average WorkKeys standard DOK level and the average WorkKeys item DOK level was 0.38.

Results by Sub-Study

The results of each sub-study are in the next sections (Sub-Study 1, NAEP Grade 12 *Reading* items to NAEP Grade 12 *Reading* standards; Sub-Study 2, WorkKeys *Reading for Information* items to NAEP Grade 12 *Reading* items; Sub-Study 3, NAEP Grade 12 *Reading* items to WorkKeys *Reading for Information* standards; and Sub-Study 4, WorkKeys *Reading for Information* items to WorkKeys *Reading for Information* standards). A summary table is presented for each sub-study, with a discussion of the results and interpretation following. Complete data and tables are available in the report appendices.

Sub-Study 1: NAEP Grade 12 *Reading* Items to NAEP Grade 12 *Reading* Standards

As described earlier, panelists have the option to code an item to a "generic" standard — the content statement at the head of a group of objectives — if they feel that the item does not clearly

align to a particular objective. In Sub-Study 1, the alignment between the NAEP Grade 12 *Reading* items and the NAEP standards, four of 131 NAEP test items (3.05%) were coded to a generic NAEP standard by at least one panelist, indicating that there was a small number of items that some panelists did not feel aligned precisely to any specific objective. There were no items the panelists deemed uncodable.

Table 9 shows a summary of the results of Sub-Study 1. The four alignment criteria analyzed are Categorical Concurrence, Depth-of-Knowledge Consistency, Range of Knowledge, and Balance of Representation. The table shows whether the two panels' judgments resulted in the four alignment criteria being met ("Yes"), weakly met ("Weak"), or not met ("No"). The degree to which the alignment criteria are met is determined by whether the calculations associated with each criterion result in values that meet predetermined threshold values that are programmed in the WAT software. These threshold values are as follows:

- For Categorical Concurrence, the threshold values used are: 6 or more for "Yes"; 5 for "Weak"; and fewer than 5 for "No."
- For Depth-of-Knowledge Consistency, the threshold values used are: 50% or more for "Yes"; 41% – 49% for "Weak"; and 40% or less for "No."
- For Range of Knowledge, the threshold values used are: 50% or more for "Yes"; 41% – 49% for "Weak"; and 40% or less for "No."
- For Balance of Representation, the threshold values used are: 0.70 – 1.0 for "Yes"; 0.61 – 0.69 for "Weak"; and 0.60 or less for "No."

Asterisks are used to denote values considered "Weak" or "No" according to the WAT threshold values. One asterisk (*) indicates that the standard **weakly** meets the alignment criterion according to the threshold values outlined above. Two asterisks (**) indicate that the standard does **not** meet the alignment criterion according to the threshold values.

Table 9: Sub-Study 1 — NAEP Grade 12* Reading *items to NAEP Grade 12* Reading *standards

NAEP Reading Standards	Sub-Study 1 — Panels 1 and 2 NAEP Grade 12 *Reading* Items Alignment Criteria							
	Categorical Concurrence (mean hits)		Depth-of-Knowledge Consistency (% of hits at or above DOK level of standard)		Range of Knowledge (% of objectives hit)		Balance of Representation (balance index)	
	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
1) Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension.	37	38	90	91	62	59	0.54**	0.52**
2) Integrate/Interpret: Make complex inferences within and across texts.	75.33	78.57	76	64	79	89	0.63*	0.64*
3) Critique/Evaluate: Consider text(s) critically.	22	17.71	88	94	72	61	0.65*	0.69*

Table 9 shows 24 points for which the degree of alignment between the NAEP items and the NAEP standards is calculated. The table shows that the panels' judgment resulted in the following:

- Alignment criteria met at 18 of 24 points (75%)
- Weak alignment at 4 of 24 points (16.67%)
- No alignment at 2 of 24 points (8.33%)

Categorical Concurrence, Depth-of-Knowledge Consistency, and Range of Knowledge:
The alignment criteria of Categorical Concurrence, Depth-of-Knowledge Consistency, and Range of Knowledge criteria were met for all three standards. In other words, there were six or more items that targeted each of the three standards, the majority of those items were at or above the DOK levels of the objectives to which they were coded, and at least 50% of all objectives within each standard were hit by at least one item.

Balance of Representation:
In contrast, the Balance of Representation criterion was not fully met across the three standards. This indicates that of the objectives that were hit (a "hit" being defined as one reviewer coding an item to an objective), the test items were not evenly distributed. Across all 37 objectives, Panel 1

had 19 objectives to which a "low" number of items or no items were coded, and Panel 2 had 18 such objectives. (Note: The standard calculations and thresholds that are programmed into the WAT are used here as the definition of "low.") In other words, Panel 1 coded a low number of items or no items to 51% of the objectives, and Panel 2 coded a low number of items or no items to 49% of the objectives.

Thus, the results suggest that the NAEP items strongly target approximately half of the objectives on which the test is based.

Looking more closely at how the NAEP items were coded to the NAEP objectives, Table 10 displays the number and percentage of mean hits to objectives. Percentages for this table are percentage of total hits.

Table 10: Number and percentage of mean hits to objectives as rated by 13 reviewers — NAEP Grade 12* Reading *items to NAEP Grade 12* Reading *standards

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
1	1.1.a	20.50	15%	24.71	18%
	1.2.a	1.17	1%	0.29	0%
	1.2.b	1.17	1%	1.71	1%
	1.2.c	0.00	0%	0.14	0%
	1.2.d	0.17	0%	0.14	0%
	1.2.e	0.00	0%	0.29	0%
	1.3 (Generic)	0.17	0%	0.29	0%
	1.3.a	5.67	4%	4.14	3%
	1.3.b	1.83	1%	2.43	2%
	1.3.c	5.67	4%	3.71	3%
	1.3.d	0.67	0%	0.14	0%
2	2.1 (Generic)	0.50	0%	0.00	0%
	2.1.a	1.00	1%	0.57	0%
	2.1.b	10.33	8%	9.29	7%
	2.1.c	1.17	1%	2.43	2%
	2.1.d	4.83	4%	4.14	3%
	2.1.e	3.83	3%	3.57	3%
	2.1.f	4.50	3%	3.57	3%
	2.2.a	1.17	1%	2.29	2%
	2.2.b	1.17	1%	1.71	1%
	2.2.c	6.00	4%	6.43	5%
	2.2.d	2.33	2%	3.86	3%
	2.2.e	0.00	0%	1.14	1%
	2.3.a	5.83	4%	2.86	2%
	2.3.b	6.17	5%	7.14	5%
	2.3.c	1.33	1%	3.14	2%
	2.3.d	0.00	0%	0.00	0%
	2.3.e	0.33	0%	1.00	1%
	2.4.a	24.83	18%	25.43	19%
3	3.1.a	6.83	5%	5.00	4%
	3.1.b	4.33	3%	2.43	2%
	3.1.c	1.33	1%	0.43	0%
	3.2.a	1.67	1%	0.71	1%
	3.2.b	0.50	0%	0.00	0%
	3.2.c	1.17	1%	2.00	1%
	3.3.a	0.83	1%	0.00	0%
	3.3.b	5.00	4%	5.57	4%
	3.3.c	0.33	0%	0.86	1%
	3.3.d	0.00	0%	0.71	1%

Both panels coded at least one NAEP item to 81% of the NAEP objectives (not including "generic" objectives). One or both panels coded no items to 19% of the objectives. The objectives to which one or both panels coded no items are as follows:

- 1.2.c — "Locate or recall setting"
- 1.2.e — "Locate or recall organizing structures of literary texts, such as verse or stanza in poetry or description, chronology, comparison, etc. in literary non-fiction"
- 2.2.e — "Explain how rhythm, rhyme, sound, or form in poetry contribute to meaning"
- 2.3.d — "Distinguish facts from opinions"
- 3.2.b — "Determine the degree to which literary devices enhance a literary work"
- 3.3.a — "Evaluate the way the author selects language to influence readers"
- 3.3.d — "Judge the coherence or logic of an argument"

Sub-Study 2: WorkKeys *Reading for Information* Items to NAEP Grade 12 *Reading* Standards

In Sub-Study 2, the alignment between the WorkKeys *Reading for Information* items and the NAEP Grade 12 *Reading* standards, there were no items coded to generic standards. That is, panelists felt that all WorkKeys test items aligned to particular NAEP objectives. In addition, there were no items that were deemed uncodable.

Table 11 shows a summary of the results of Sub-Study 2. The four alignment criteria analyzed are Categorical Concurrence, Depth-of-Knowledge Consistency, Range of Knowledge, and Balance of Representation. The table shows whether the two panels' judgments resulted in the four alignment criteria being met ("Yes"), weakly met ("Weak"), or not met ("No"). The degree to which the alignment criteria are met is determined by whether the calculations associated with each criterion result in values that meet predetermined threshold values that are programmed in the WAT software. These threshold values are as follows:

- For Categorical Concurrence, the threshold values used are: 6 or more for "Yes"; 5 for "Weak"; and fewer than 5 for "No."
- For Depth-of-Knowledge Consistency, the threshold values used are: 50% or more for "Yes"; 41% – 49% for "Weak"; and 40% or less for "No."
- For Range of Knowledge, the threshold values used are: 50% or more for "Yes"; 41% – 49% for "Weak"; and 40% or less for "No."
- For Balance of Representation, the threshold values used are: 0.70 – 1.0 for "Yes"; 0.61 – 0.69 for "Weak"; and 0.60 or less for "No."

Asterisks are used to denote values considered "Weak" or "No" according to the WAT threshold values. One asterisk (*) indicates that the standard **weakly** meets the alignment criterion according to the threshold values outlined above. Two asterisks (**) indicate that the standard does **not** meet the alignment criterion according to the threshold values.

Table 11: Sub-Study 2 — WorkKeys* Reading for Information *items to NAEP Grade 12* Reading *standards

NAEP Reading Standards	Sub-Study 2 — Panels 1[1] and 2 WorkKeys *Reading for Information* Items Alignment Criteria							
	Categorical Concurrence (mean hits)		Depth-of-Knowledge Consistency (% of hits at or above DOK level of standard)		Range of Knowledge (% of objectives hit)		Balance of Representation (balance index)	
	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
1) Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension.	42.4	42.71	80	76	46*	40**	0.64*	0.68*
2) Integrate/Interpret: Make complex inferences within and across texts.	18	17.29	53	46*	35**	40*	0.71	0.70
3) Critique/Evaluate: Consider text(s) critically.	0**	0**	0**	0**	0**	0**	0**	0**

[1] Panel 1 had five of the six members during this sub-study because one panelist was ill.

Table 11 shows 24 points for which the degree of alignment between the WorkKeys *Reading for Information* items and the NAEP reading standards is calculated. The table shows that the panels' judgment resulted in the following:

- Alignment criteria met at 9 of 24 points (37.50%)
- Weak alignment at 5 of 24 points (20.83%)
- No alignment at 10 of 24 points (41.67%)

The data show some degree of alignment for Standard 1, Locate/Recall, and for Standard 2, Integrate/Interpret, but not for Standard 3, Critique/Evaluate.

Similarly, the alignment criteria of Categorical Concurrence and Depth-of-Knowledge Consistency were met for Standard 1, Locate/Recall, and for Standard 2, Integrate/Interpret, but not for Standard 3, Critique/Evaluate.

Categorical Concurrence:
The Categorical Concurrence criterion was met for Standards 1 and 2. These standards require individuals to "Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension"

(Standard 1) and to "Integrate/Interpret: Make complex inferences within and across texts" (Standard 2). As may be expected, the objectives within each standard that received the most hits were those focused on informational text, as opposed to literary text.

In contrast, no WorkKeys items aligned to the objectives within Standard 3, "Critique/Evaluate: Consider text(s) critically." A significant reason for this is the fact that all WorkKeys items are multiple choice. The panelists felt that it would be highly unusual for a multiple-choice item to assess the cognitive skill of critiquing/evaluating, and they did not see evidence of this type of assessment in the WorkKeys item pool used for this study.

Depth-of-Knowledge Consistency:
One minor point of difference between Panels 1 and 2 was that Panel 2 tended to rate the DOK levels for the WorkKeys items lower than the DOK levels for the NAEP objectives to which they were aligned, whereas Panel 1 did not. The threshold established in the WAT for Depth-of-Knowledge Consistency is that at least 50% of the items must be coded at or above the DOK level of the standard in order for the criterion to be considered to be met. Panel 2 members assigned lower DOK levels to the WorkKeys items than to the Standard 2 objectives to which they were aligned more frequently than did Panel 1 members.

Looking more closely at how the WorkKeys items were coded to the NAEP objectives, Table 12 displays the number and percentage of mean hits to objectives.

Table 12: Number and percentage of mean hits to objectives as rated by 12 reviewers — WorkKeys* Reading for Information *Items to NAEP Grade 12* Reading *standards

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
1	1.1.a	24.00	40%	24.43	41%
	1.2.a	0.20	0%	0.00	0%
	1.2.b	6.60	11%	7.14	12%
	1.2.c	0.20	0%	0.00	0%
	1.2.d	0.00	0%	0.00	0%
	1.2.e	0.00	0%	0.00	0%
	1.3.a	0.20	0%	0.00	0%
	1.3.b	2.00	3%	2.14	4%
	1.3.c	9.20	15%	9.00	15%
	1.3.d	0.00	0%	0.00	0%
2	2.1.a	1.80	3%	1.14	2%
	2.1.b	4.60	8%	4.00	7%
	2.1.c	0.00	0%	1.14	2%
	2.1.d	0.00	0%	0.00	0%
	2.1.e	0.00	0%	0.00	0%
	2.1.f	1.00	2%	1.00	2%
	2.2.a	0.00	0%	0.00	0%
	2.2.b	0.00	0%	0.00	0%
	2.2.c	0.00	0%	0.00	0%
	2.2.d	0.00	0%	0.00	0%
	2.2.e	0.00	0%	0.00	0%
	2.3.a	1.20	2%	1.00	2%
	2.3.b	4.60	8%	4.00	7%
	2.3.c	0.00	0%	0.14	0%
	2.3.d	0.00	0%	0.00	0%
	2.3.e	0.00	0%	0.14	0%
	2.4.a	4.80	8%	4.71	8%
3	3.1.a	0.00	0%	0.00	0%
	3.1.b	0.00	0%	0.00	0%
	3.1.c	0.00	0%	0.00	0%
	3.2.a	0.00	0%	0.00	0%
	3.2.b	0.00	0%	0.00	0%
	3.2.c	0.00	0%	0.00	0%
	3.3.a	0.00	0%	0.00	0%
	3.3.b	0.00	0%	0.00	0%
	3.3.c	0.00	0%	0.00	0%
	3.3.d	0.00	0%	0.00	0%

Range of Knowledge:

Table 12 further illustrates why the Range of Knowledge alignment criterion was met either weakly or not at all for all three NAEP standards. An examination of the data shows that, for

Standard 1, Panel 1 determined that three of the ten NAEP objectives were not targeted by a WorkKeys item. Panel 2 determined that six of the ten objectives were not targeted. Objectives not targeted are related to features of literary texts, such as setting, figurative language, and organizing structures; thus, it is not surprising that this criterion is not met, given the workplace, non-literary orientation of the WorkKeys assessment reading passages.

It is important to note here that for Panel 2, the Range of Knowledge value is 40 for both Standard 1 and Standard 2. Typically, this would result in an indication of "no alignment" for both standards. However, the WAT software indicates "no alignment" for Standard 1 but "weak alignment" for Standard 2. The reason for this lies in the standard deviation. The standard deviation is 0 for Standard 1, but it is 4 for Standard 2. There was strong agreement among panelists on the degree of alignment within Standard 1, but greater variety of results among panelists for Standard 2. The nature of these differences was such that there were more indications of alignment for Standard 2 than for Standard 1.

Balance of Representation:
Table 12 also illustrates how the Balance of Representation alignment criterion was weakly met for Standard 1, Locate/Recall, meaning that there was some tendency for items to cluster on one objective. As an example, 26 of the 60 WorkKeys items targeted objective 1.1.a ("Locate or recall specific information such as definition, facts, and supporting details in text or graphics").

The Balance of Representation criterion was met ("Yes") for Standard 2, Integrate/Interpret. Even so, no objectives under 2.2 ("Integrate/Interpret: Make complex inferences within and across literary texts") were targeted. Other objectives not targeted were related to author's technique, literary devices, organizing structures, separating fact from opinion, and identifying tacit assumptions. These findings result in large part from the fact that the WorkKeys test does not include any reading passages of a literary nature, whereas a significant portion of the NAEP standards is focused on literary content.

As described previously, no WorkKeys items targeted any objectives under Standard 3, Critique/Evaluate.

Both panels coded at least one WorkKeys item to 27% of the NAEP objectives (not including "generic" objectives). One or both panels coded no items to 83% of the objectives. The objectives to which one or both panels coded no items are as follows:

- 1.2.a — "Locate or recall character traits"
- 1.2.c — "Locate or recall setting"
- 1.2.d — "Locate or recall figurative language"
- 1.2.e — "Locate or recall organizing structures of literary texts, such as verse or stanza in poetry or description, chronology, comparison, etc. in literary non-fiction"
- 1.3.a — "Locate or recall the topic sentence or main idea"
- 1.3.d — "Locate or recall organizing structures of texts, such as comparison/contrast, problem/solution, enumeration, etc."
- 2.1.c — "Determine unstated assumptions in an argument"
- 2.1.d — "Describe or analyze how an author uses literary devices or text features to convey meaning"

- 2.1.e — "Describe or analyze how an author uses organizing structures to convey meaning"
- 2.2.a — "Interpret mood, tone, or voice"
- 2.2.b — "Integrate ideas to determine theme"
- 2.2.c — "Interpret a character's conflicts, motivations, and decisions"
- 2.2.d — "Examine relations between or among theme, setting, plot, or characters"
- 2.2.e — "Explain how rhythm, rhyme, sound, or form in poetry contribute to meaning"
- 2.3.c — "Find evidence in support of an argument"
- 2.3.d — "Distinguish facts from opinions"
- 2.3.e — "Determine the importance of information within and across texts"
- All of the objectives within Standard 3, "Critique/Evaluate: Consider text(s) critically."

The NAEP objectives to which no WorkKeys items aligned include those specific to literary text and literary devices, and all of the objectives under Standard 3, "Critique/Evaluate: Consider text(s) critically."

Sub-Study 3: NAEP Grade 12 *Reading* Items to WorkKeys *Reading for Information* Standards

In Sub-Study 3, the alignment between the NAEP Grade 12 *Reading* items and the WorkKeys *Reading for Information* standards, 73 of 131 NAEP items (55.73%) were rated as uncodable to WorkKeys standards by at least one panelist. Of these, 10 items (7.63%) were deemed uncodable by all panelists. In addition, 16 NAEP items (12.21%) were coded to a generic WorkKeys standard by at least one rater, indicating that there was a small number of items that some panelists did not feel aligned precisely to any specific objective. One NAEP item (0.76%) was coded to a generic standard by all panelists.

Table 13 shows a summary of the results of Sub-Study 3. The four alignment criteria analyzed are Categorical Concurrence, Depth-of-Knowledge Consistency, Range of Knowledge, and Balance of Representation. The table shows whether the two panels' judgments resulted in the four alignment criteria being met ("Yes"), weakly met ("Weak"), or not met ("No"). The degree to which the alignment criteria are met is determined by whether the calculations associated with each criterion result in values that meet predetermined threshold values that are programmed in the WAT software. These threshold values are as follows:

- For Categorical Concurrence, the threshold values used are: 6 or more for "Yes"; 5 for "Weak"; and fewer than 5 for "No."
- For Depth-of-Knowledge Consistency, the threshold values used are: 50% or more for "Yes"; 41% – 49% for "Weak"; and 40% or less for "No."
- For Range of Knowledge, the threshold values used are: 50% or more for "Yes"; 41% – 49% for "Weak"; and 40% or less for "No."
- For Balance of Representation, the threshold values used are: 0.70 – 1.0 for "Yes"; 0.61 – 0.69 for "Weak"; and 0.60 or less for "No."

Asterisks are used to denote values considered "Weak" or "No" according to the WAT threshold values. One asterisk (*) indicates that the standard **weakly** meets the alignment criterion

according to the threshold values outlined above. Two asterisks (**) indicate that the standard does **not** meet the alignment criterion according to the threshold values.

Table 13: Sub-Study 3 — NAEP Grade 12* Reading *items to WorkKeys* Reading for Information *standards

WorkKeys *Reading for Information* Standards	Sub-Study 3 — Panels 1 and 2 NAEP Grade 12 *Reading* Items Alignment Criteria							
	Categorical Concurrence (mean hits)		Depth-of-Knowledge Consistency (% of hits at or above DOK level of standard)		Range of Knowledge (% of objectives hit)		Balance of Representation (balance index)	
	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
3) Individuals read short, simple, and clearly stated materials to find out what should be done.	27	27.14	69	33**	43*	23**	0.65*	0.94
4) Individuals read straightforward information that contains a number of details. When following procedures, they must think about changing conditions that affect what should be done.	5.33*	6.29	55	75	50	43*	0.84	0.88
5) Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions.	10.67	13.57	78	93	17**	21**	1	0.88

WorkKeys *Reading for Information* Standards	Sub-Study 3 — Panels 1 and 2 NAEP Grade 12 *Reading* Items Alignment Criteria							
	Categorical Concurrence (mean hits)		Depth-of-Knowledge Consistency (% of hits at or above DOK level of standard)		Range of Knowledge (% of objectives hit)		Balance of Representation (balance index)	
	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
6) Individuals read elaborate procedures, complicated information, and legal regulations, all of which contain difficult words, jargon, and technical terms. Most information is not clearly stated.	35	33.71	68	75	52	49*	0.68*	0.71
7) Individuals read very complex information which includes a lot of details and complicated concepts. Unusual jargon and technical terms are used but not defined. Writing often lacks clarity and direction. Individuals must draw conclusions from some parts of the reading and apply them to other parts.	13.5	9.57	86	100	54	50	0.85	0.61*

Table 13 shows 40 points for which the degree of alignment between the NAEP items and the WorkKeys standards is calculated. The table shows that the panels' judgment resulted in the following:

- Alignment criteria met at 29 of 40 points (72.5%)
- Weak alignment at 7 of 40 points (17.5%)
- No alignment at 4 of 40 points (10%)

For the alignment criteria of Categorical Concurrence and Depth-of-Knowledge Consistency, the data show alignment within both panels, with two exceptions.

Categorical Concurrence:
The exception for the criterion of Categorical Concurrence is found at Standard 4 for Panel 1. Here, the mean hits value is 5.33, just below the 6-item threshold for categorical concurrence. Thus, the judgments of the two panels show that there are at least 6 NAEP items targeting at least one WorkKeys objective within each standard except Standard 4, which reads as follows: "Individuals read straightforward information that contains a number of details. When following procedures, they must think about changing conditions that affect what should be done."

One additional detail in particular related to the alignment for Standard 7 was discussed by panelists and is worth noting here. WorkKeys generic Standard 7 is as follows: "Individuals read very complex information which includes a lot of details and complicated concepts. Unusual jargon and technical terms are used but not defined. Writing often lacks clarity and direction. Individuals must draw conclusions from some parts of the reading and apply them to other parts." As explained earlier in this report, the Webb methodology encourages panelists to code items to the specific objectives beneath a given standard whenever possible. In Sub-Study 3, panelists did not code to any generic standards besides Standard 7. Categorical Concurrence alignment to this standard is due almost entirely to items being coded to the generic standard or to Objective 7.1 ("Figure out the definitions of difficult, uncommon words based on how they are used"). Thus, although the calculations made by the WAT indicate that there is alignment here, it is important to note that the panelists did not feel there was an appropriate specific objective to which to align NAEP items that otherwise seemed to fit at WorkKeys Level 7.

A reason for this characteristic of the panelists' coding at Level 7 of the WorkKeys standards is likely due to a difference in the nature of NAEP and WorkKeys reading passages, and the panelists' comments recorded in the WAT for this sub-study reflect this. One panelist noted, "Many NAEP items assessed aspects related to literary text and author's craft. Also, the NAEP items included those asking for a critique/evaluation, which the WorkKeys did not include. The WorkKeys objectives, on the other hand, emphasized instructions, policies, procedures, and jargon, which the NAEP items did not assess." Another panelist wrote, "…it is difficult to assign standards based on certain types of informational text to passages that are literary or narrative in scope. Additionally, the standards are focused on more business-oriented texts (e.g., manuals and memos) which is not illustrated on the NAEP. Levels of difficulty are quite distinct for each assessment and it was as if you were to compare apples to oranges. NAEP passages, by and large, are well-written … and not considered to be 'functional texts' beyond perceived literary pleasure or gaining knowledge about world events or perspectives. NAEP passages are also longer in scope and affiliated test items look for application via written support and explanations. Such requests are not reflected in the WorkKeys objectives." The passages on the NAEP assessment are typically well written. In contrast, the WorkKeys assessment is designed to reflect workplace realism, and the reading passages are taken from real-world samples and, therefore, reflect the varied writing abilities of the individuals who produced them. Particularly at the higher levels, such as Level 7, the passages are complex both in terms of content and writing style, and because they may lack clarity and direction. This difference also contributed to some of the challenge the panelists experienced as they coded the NAEP items to the WorkKeys standards.

Further insight into the alignment between the NAEP items and the WorkKeys standards using the Categorical Concurrence criterion may be gained by considering the items that were deemed uncodable. Sub-Study 3 was the only sub-study for which there were items deemed uncodable by all panelists. Ten items were not coded by anyone in either group. The following tables address the issue of uncodable items.

Table 14 displays the counts of items determined to be codable and uncodable by all raters in a panel. Each item is counted once and totals are not weighted by point value.

Table 14: Codability of items as determined by items rated uncodable by 100% of reviewers — NAEP Grade 12* Reading *items to WorkKeys* Reading for Information *standards

	Panel 1	Panel 2
Codable items	111	107
Uncodable items	20	24
Total assessment items	131	131

Table 15 displays the distribution of panelist item ratings by codable and uncodable. All items are weighted equally, and the mean codable items are calculated by dividing the number of item ratings by the number of reviewers.

Table 15: Number and percentage of mean hits (codable and uncodable) as rated by 13 reviewers — NAEP Grade 12* Reading *items to WorkKeys* Reading for Information *standards

	Panel 1		Panel 2	
	Mean Hits	Percentage	Mean Hits	Percentage
Codable	91.5	69.85	90.29	68.92
Uncodable	39.5	30.15	40.71	31.08
Total	131		131	

Table 16 displays the categorical concurrence and distribution of panelist item ratings across the standards. Percentage of hits is presented in two ways: 1) as the percentage of codable items; and 2) as adjusted percentages to include all items, codable and uncodable.

Table 16: Categorical concurrence between standards and assessment as rated by 13 reviewers — NAEP Grade 12* Reading *items to WorkKeys* Reading for Information *standards

	Panel 1			Panel 2		
WorkKeys Standard	Mean Hits	% of Codable Hits	% Hits, Adjusted for Uncodable	Mean Hits*	% of Codable Hits	% Hits, Adjusted for Uncodable
3	27	29.51%	20.61%	27.14	30.06%	20.72%
4	5.33	5.83%	4.07%	6.29	6.97%	4.80%
5	10.67	11.66%	8.14%	13.57	15.03%	10.36%
6	35	38.25%	26.72%	33.71	37.34%	25.73%
7	13.5	14.75%	10.31%	9.57	10.6%	7.31%
Total	91.5	100%	69.85%	90.29	100%	68.92%

**These numbers are taken directly from the WAT-generated reports. The minor discrepancy in summing this column is hypothesized to be due to rounding in the WAT calculations.*

Thus, it can be seen that the panelists found roughly 70% of the NAEP items to be codable to the WorkKeys standards. Furthermore, Standard 6 received the greatest percentage of hits, followed by Standards 3, then 7 and 5, then 4.

Depth-of-Knowledge Consistency:
The data show that the alignment criterion of Depth-of-Knowledge Consistency was met in the judgment of both panels at all standards, with one exception. The Standard 3 data from Panel 2 show that those panelists judged the NAEP item DOK levels at or above the DOK levels of the WorkKeys objective to which they were coded only 33% of the time. Examination of the data shows that Panel 2 coded NAEP items to only one objective within Standard 3, Objective 3.5. This objective has a DOK level of 2, while the majority of items coded to this objective have a DOK level of 1. Panel 1 had at least one rater that assigned items with DOK levels of 1 to other objectives with a DOK level of 1 within the standard. For this panel, the addition of one or two raters assigning items with a DOK level of 1 to objectives with a DOK level of 1 was sufficient to yield different results from those of Panel 2.

In contrast to Categorical Concurrence and Depth-of-Knowledge Consistency, the NAEP items aligned less strongly to the WorkKeys standards using the Range of Knowledge and Balance of Representation alignment criteria, the two criteria related to how the aligned items are distributed among the objectives within a standard.

Range of Knowledge:
The NAEP items and WorkKeys standards did not meet the threshold values for the Range of Knowledge alignment criterion at Standard 3 ("Individuals read short, simple, and clearly stated materials to find out what should be done") or at Standard 5 ("Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions"). In addition, the data for Panel 2 weakly meet the threshold values for this criterion at Standards 4 and 6. And taking the results as a whole, neither panel judged the alignment between the NAEP items and the WorkKeys standards to be greater than 54% for any of the standards.

Standard 3: This standard reads, "Individuals read short, simple, and clearly stated materials to find out what should be done." The coding of the two panels resulted in differing conclusions about Range of Knowledge for Standard 3. Panel 1 results show that there is weak alignment, while Panel 2 results show that there is not alignment.

If the Range of Knowledge criterion is met, it means that at least half of the objectives within the standard are targeted by at least one item. Panel 1 aligned items with all five objectives under Standard 3. A notable feature of Panel 1 coding is that one panelist coded a number of NAEP items to WorkKeys Objective 3.1 ("Apply instructions to a situation that is the same as the one in the reading materials"), and the rest of the panelists aligned those items to Objective 3.5 ("Identify main ideas and clearly stated details"). One Panel 1 member coded two items to Objective 3.2 ("Choose the correct meaning of a word that is clearly defined in the reading"), and another Panel 1 member coded an item to Objective 3.4 ("Choose when to perform each step in a short series of steps"). Thus, the WAT calculations show that Panel 1 deemed the NAEP items meet this alignment criterion, as 43% of the objectives for the standard were hit (at least 40% of the objectives must receive at least one hit in order to meet the threshold for "weak" alignment).

In contrast to Panel 1, Panel 2 members targeted only one objective within Standard 3 — 3.5 ("Identify main ideas and clearly stated details") — *except* that one panelist coded one item to

Objective 3.3 ("Choose the correct meaning of common, everyday workplace words"). Because all but one of the items that were coded to Standard 3 were coded to just one of five possible objectives, the Range of Knowledge criterion was not met for Panel 2 — just 23% of the objectives were hit.

Standard 5: The data from both panels do not meet the threshold values for alignment using the Range of Knowledge criterion. Examination of the data shows that NAEP items cluster on objective 5.4 ("Figure out the correct meaning of a word based on how the word is used"). Objectives under Standard 5 that were *not* targeted are related to following instructions that have conditional statements; applying instructions to similar situations; and word definitions regarding technical term, jargon, and acronyms. Panelists established decision rules related to some of the items and standards discussed at this point, including Decision Rules 7 and 11.

Decision Rule 7 establishes that NAEP items about topics such as literary devices and author's craft, and items requiring examinees to construct a response explaining or evaluating something are uncodable. However, only one panel agreed to adopt the fourth point under Decision Rule 7: "The following types of NAEP items are regarded by the panelists as uncodable to WorkKeys standards: … NAEP vocabulary items with a DOK level of 1 and associated with a literary stimulus passage." The other panel determined that such NAEP items *could* be coded to a WorkKeys standard.

Decision Rule 11 establishes how to code items associated with a reading passage that contains instructions. It was challenging for panelists to apply this rule consistently. As noted earlier in this report, "(b)oth panels struggled with this issue and found it difficult to apply the decision rule consistently. Despite having the decision rule, both panels found it necessary to adjudicate the coding for some items that fell in this category." All of these factors likely contributed to the lower degree of alignment to this standard, as well as to Standard 3.

Balance of Representation:
In the alignment methodology used for this study, the following definition of Balance of Representation is given: "An index is used to judge the distribution of assessment items among subcategories [objectives] underlying a content category [standard]. An index value of 1 signifies perfect balance and is obtained if the corresponding items related to a content category are equally distributed among the course-level expectations for the category." Thus, if calculations performed by the WAT software indicate that there is alignment according to the Balance of Representation criterion, it might be expected that the items aligned to the standard are spread among all the targeted objectives within the standard, not clustered on a small number of targeted objectives. Further discussion with Dr. Webb during the course of the data analysis phase of this research, however, clarified that, in fact, the calculation is completed *only* on the basis of the objectives to which any items are coded, not on the basis of all objectives within a given standard.

This clarification is critical in interpreting the data for this portion of Sub-Study 3.

The WAT calculations indicate Balance of Representation alignment — with index values ranging from 0.84 to 1.0 — for six of the ten points in the sub-study. The other four points show Balance of Representation alignment index values of 0.61 to 0.71. It might be inferred from these data that the NAEP items are coded to the WorkKeys standards fairly evenly. This is not the case,

however, and the following table helps to illustrate this. It shows the number and percentage of mean hits to objectives.

Table 17: Number and percentage of mean hits to objectives as rated by 13 reviewers — NAEP Grade 12* Reading *items to WorkKeys* Reading for Information *standards

WorkKeys Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
3	3.1	2.17	2%	0.00	0%
	3.2	0.33	0%	0.00	0%
	3.3	1.17	1%	0.14	0%
	3.4	0.17	0%	0.00	0%
	3.5	23.17	25%	27.00	30%
4	4.1	0.00	0%	0.00	0%
	4.2	0.00	0%	0.43	0%
	4.3	2.33	3%	3.14	3%
	4.4	3.00	3%	2.71	3%
5	5.1	0.00	0%	0.00	0%
	5.2	0.00	0%	0.14	0%
	5.3	0.00	0%	0.00	0%
	5.4	10.67	12%	13.29	15%
	5.5	0.00	0%	0.14	0%
	5.6	0.00	0%	0.00	0%
6	6.1	0.00	0%	0.00	0%
	6.2	0.33	0%	0.00	0%
	6.3	15.17	17%	14.29	16%
	6.4	3.00	3%	3.29	4%
	6.5	0.50	1%	0.43	0%
	6.6	16.00	17%	15.71	17%
	6.7	0.00	0%	0.00	0%
7	7	5.00	5%	1.00	1%
	7.1	8.33	9%	8.57	9%
	7.2	0.17	0%	0.00	0%
	7.3	0.00	0%	0.00	0%

Thus, it is possible to see that the WAT calculations consider only the objectives to which at least one item has been coded. For instance, at Standard 4, approximately 2% of the NAEP items were coded to each of Objectives 4.3 and 4.4. Given that the distribution of items among these two objectives was fairly even, the Balance of Representation index value is high (0.84 and 0.88).

Both panels coded at least one NAEP item to 40% of the WorkKeys objectives. One or both panels coded no items to 60% of the objectives. The objectives to which one or both panels coded no items are as follows:

- 3.1 — "Apply instructions to a situation that is the same as the one in the reading materials"
- 3.2 — "Choose the correct meaning of a word that is clearly defined in the reading"
- 3.4 — "Choose when to perform each step in a short series of steps"

- 4.1 — "Apply instructions with several steps to a situation that is the same as the situation in the reading materials"
- 4.2 — "Choose what to do when changing conditions call for a different action (follow directions that include 'if-then' statements)"
- 5.1 — "Apply complex instructions that include conditionals to situations described in the materials"
- 5.2 — "Apply straightforward instructions to a new situation that is similar to the one described in the material"
- 5.3 — "Apply technical terms and jargon and relate them to stated situations"
- 5.5 — "Identify the correct meaning of an acronym that is defined in the document"
- 5.6 — "Identify the paraphrased definition of a technical term or jargon that is defined in the document"
- 6.1 — "Apply complicated instructions to new situations"
- 6.2 — "Apply general principles behind policies, rules, and procedures"
- 6.7 — "Use technical terms and jargon in new situations"
- 7.2 — "Figure out the general principles behind policies and apply them to situations that are quite different from any described in the materials"
- 7.3 — "Figure out the meaning of jargon or technical terms based on how they are used"

The WorkKeys objectives to which no NAEP items aligned are related to applying instructions, understanding jargon, or understanding the general principles behind policies, rules, and procedures.

Sub-Study 4: WorkKeys *Reading for Information* Items to WorkKeys *Reading for Information* Standards

In Sub-Study 4, the alignment between the WorkKeys *Reading for Information* items and the WorkKeys *Reading for Information* standards, three of 60 WorkKeys items (5%) were marked as uncodable by one panelist each. No items were judged as uncodable by all panelists. In addition, there was one item (1.67%) that was aligned to a generic standard by all panelists.

Table 18 shows a summary of the results of Sub-Study 4. The four alignment criteria analyzed are Categorical Concurrence, Depth-of-Knowledge Consistency, Range of Knowledge, and Balance of Representation. The table shows whether the two panels' judgments resulted in the four alignment criteria being met ("Yes"), weakly met ("Weak"), or not met ("No"). The degree to which the alignment criteria are met is determined by whether the calculations associated with each criterion result in values that meet predetermined threshold values that are programmed in the WAT software. These threshold values are as follows:

- For Categorical Concurrence, the threshold values used are: 6 or more for "Yes"; 5 for "Weak"; and fewer than 5 for "No."
- For Depth-of-Knowledge Consistency, the threshold values used are: 50% or more for "Yes"; 41% – 49% for "Weak"; and 40% or less for "No."
- For Range of Knowledge, the threshold values used are: 50% or more for "Yes"; 41% – 49% for "Weak"; and 40% or less for "No."

- For Balance of Representation, the threshold values used are: 0.70 – 1.0 for "Yes"; 0.61 – 0.69 for "Weak"; and 0.60 or less for "No."

Asterisks are used to denote values considered "Weak" or "No" according to the WAT threshold values. One asterisk (*) indicates that the standard **weakly** meets the alignment criterion according to the threshold values outlined above. Two asterisks (**) indicate that the standard does **not** meet the alignment criterion according to the threshold values.

Table 18: Sub-study 4 — WorkKeys* Reading for Information *items to WorkKeys* Reading for Information *standards

WorkKeys *Reading for Information* Standards	Sub-Study 4 — Panels 1 and 2 WorkKeys *Reading for Information* Items Alignment Criteria							
	Categorical Concurrence (mean hits)		Depth-of-Knowledge Consistency (% of hits at or above DOK level of standard)		Range of Knowledge (% of objectives hit)		Balance of Representation (balance index)	
	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
3) Individuals read short, simple, and clearly stated materials to find out what should be done.	14.33	17	74	77	73	80	0.69*	0.60*
4) Individuals read straightforward information that contains a number of details. When following procedures, they must think about changing conditions that affect what should be done.	12	11.71	70	66	83	75	0.71	0.74
5) Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions.	4.67**	4.71**	88	87	36**	40*	0.72	0.75

WorkKeys *Reading for Information* Standards	Sub-Study 4 — Panels 1 and 2 WorkKeys *Reading for Information* Items Alignment Criteria							
	Categorical Concurrence (mean hits)		Depth-of-Knowledge Consistency (% of hits at or above DOK level of standard)		Range of Knowledge (% of objectives hit)		Balance of Representation (balance index)	
	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2	Panel 1	Panel 2
6) Individuals read elaborate procedures, complicated information, and legal regulations, all of which contain difficult words, jargon, and technical terms. Most information is not clearly stated.	25.83	23.14	49*	36**	86	84	0.71	0.73
7) Individuals read very complex information which includes a lot of details and complicated concepts. Unusual jargon and technical terms are used but not defined. Writing often lacks clarity and direction. Individuals must draw conclusions from some parts of the reading and apply them to other parts.	3.17**	3**	69	50	54	50	0.85	0.83

Table 18 shows 40 points for which the degree of alignment between the WorkKeys items and the WorkKeys standards is calculated. The table shows that the panels' judgment resulted in the following:

- Alignment criteria met at 30 of 40 points (75%)
- Weak alignment at 4 of 40 points (10%)
- No alignment at 6 of 40 points (15%)

Categorical Concurrence:
This criterion was met for Standards 3, 4, and 6, but not for Standards 5 and 7. The threshold set in the WAT software for this criterion is that at least six items must be coded to objectives within a given standard in order for the criterion to be considered to be met.

Standard 5 states, "Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions." Only four items aligned to this standard, according to a majority of panelists. Furthermore, only three of the six objectives were targeted by

an item Objectives to which panelists did not code any items included 5.2 ("Apply straightforward instructions to a new situation that is similar to the one described in the material") and 5.3 ("Apply technical terms and jargon and relate them to stated situations"). Few panelists coded items to Objective 5.4 ("Figure out the correct meaning of a word based on how the word is used") or to 5.6 ("Identify the paraphrased definition of a technical term or jargon that is defined in the document"). An examination of their comments revealed that several panelists believed that, while the characteristics of the text might match a higher standard, the items seemed to target lower objectives.

Standard 7 states, "Individuals read very complex information which includes a lot of details and complicated concepts. Unusual jargon and technical terms are used but not defined. Writing often lacks clarity and direction. Individuals must draw conclusions from some parts of the reading and apply them to other parts." Panelists coded three items to this standard. One item was coded to the generic standard and the other two items aligned with Objective 7.3 ("Figure out the meaning of jargon or technical terms based on how they are used"). One panelist aligned an item to Objective 7.1 ("Figure out the definitions of difficult, uncommon words based on how they are used"). Because there were not at least six items aligned to objectives within this standard, the WAT calculations showed that the Categorical Concurrence alignment criterion was not met.

Depth-of-Knowledge Consistency:
There was alignment using the Depth-of-Knowledge Consistency criterion for all standards except 6, which states, "Individuals read elaborate procedures, complicated information, and legal regulations, all of which contain difficult words, jargon, and technical terms. Most information is not clearly stated." Panelists rated the majority of these items at a lower DOK than the standard. For Standard 6, the consensus among the panelists was that four objectives were DOK Level 2, and three objectives were DOK Level 3. However, the items were coded primarily as DOK Level 2.

Range of Knowledge:
The Range of Knowledge criterion was met for all standards except Standard 5. Panel 1 found this alignment criterion to be not met, while Panel 2 found this alignment to be weak. Standard 5 states, "Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions." Both panels coded items to three of the six objectives under this standard; the WAT calculations require that *more* than half of the objectives have items coded to them in order for the criterion to be considered met.

Balance of Representation:
All standards met the criterion for alignment in the Balance of Representation category. However, the alignment to Standard 3 was weak. Standard 3 states, "Individuals read short, simple, and clearly stated materials to find out what should be done." Most of the items that were coded to Standard 3 aligned with Objective 3.5 ("Identify main ideas and clearly stated details"). Because the items at this standard were clustered at one objective, the balance of representation was weak in both panels. Panelists did not code any items to Objective 3.2 ("Choose the correct meaning of a word that is clearly defined in the reading").

That some of the objectives were not targeted by items might be a result of item sampling. These studies used two intact WorkKeys test forms. Each form has a specified number of items at each standard, the items sample a range of objectives within each standard, and the items are set within contexts that reflect the variety of careers defined by ACT's World-of-Work Map (http://www.act.org/wwm/). While all of the 131 items available on the NAEP assessment were studied, only 60 unique items from the complete pool of hundreds of operational WorkKeys items were used in these studies. In addition, the WorkKeys framework has five standards, whereas the NAEP framework has three. Both of these factors — fewer WorkKeys items and more WorkKeys standards — combine to make it more difficult for all WorkKeys objectives to be targeted by the items in a given WorkKeys test form. It is possible that different forms of the WorkKeys tests would have items targeting other objectives within the standards. However, although a different form might show alignment to different objectives within a standard, thereby altering the range and balance criteria, it is unlikely that another form would yield a different overall alignment to the standards because WorkKeys test forms are equated and parallel.

Looking more closely at how the WorkKeys items were coded to the WorkKeys objectives, Table 19 displays the number and percentage of mean hits to objectives.

Table 19: Number and percentage of mean hits to objectives as rated by 13 reviewers — WorkKeys* Reading for Information *items to WorkKeys* Reading for Information *standards

Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
3	3.1	3.50	6%	3.00	5%
	3.2	0.00	0%	0.00	0%
	3.3	0.67	1%	1.00	2%
	3.4	2.00	3%	2.00	3%
	3.5	8.17	14%	11.00	18%
4	4.1	2.67	4%	2.43	4%
	4.2	2.00	3%	2.29	4%
	4.3	7.00	12%	7.00	12%
	4.4	0.33	1%	0.00	0%
5	5.1	3.50	6%	3.14	5%
	5.2	0.00	0%	0.00	0%
	5.3	0.00	0%	0.00	0%
	5.4	0.00	0%	0.57	1%
	5.5	1.00	2%	1.00	2%
	5.6	0.17	0%	0.00	0%
6	6.1	2.17	4%	2.00	3%
	6.2	7.50	13%	7.43	12%
	6.3	4.83	8%	4.86	8%
	6.4	2.83	5%	2.43	4%
	6.5	1.33	2%	1.29	2%
	6.6	7.17	12%	5.14	9%
	6.7	0.00	0%	0.00	0%
7	7 (Generic)	1.00	2%	1.00	2%
	7.1	0.17	0%	0.00	0%
	7.2	0.00	0%	0.00	0%
	7.3	2.00	3%	2.00	3%

Both panels coded at least one WorkKeys item to 64% of the WorkKeys objectives (not including "generic" objectives). One or both panels coded no items to 36% of the objectives. The objectives to which one or both panels coded no items are as follows:

- 3.2 — "Choose the correct meaning of a word that is clearly defined in the reading"
- 4.4 — "Use the reading material to figure out the meaning of words that are not defined"
- 5.2 — "Apply straightforward instructions to a new situation that is similar to the one described in the material"
- 5.3 — "Apply technical terms and jargon and relate them to stated situations"
- 5.4 — "Figure out the correct meaning of a word based on how the word is used"
- 5.6 — "Identify the paraphrased definition of a technical term or jargon that is defined in the document"
- 6.7 — "Use technical terms and jargon in new situations"
- 7.1 — "Figure out the definitions of difficult, uncommon words based on how they are used"

- 7.2 — "Figure out the general principles behind policies and apply them to situations that are quite different from any described in the materials"

Most of the WorkKeys objectives to which one or both panels did not code WorkKeys items are related to vocabulary.

Panelist Evaluation Results

The panelists completed evaluation surveys after each main task of the alignment study. Following is a summary of their responses to each. The full compilation of responses is in Appendix G.

Training Questionnaire

Panelists were presented with six questions about the effectiveness of the training presented on Day 1, for which the possible responses were Not Well (1) Somewhat (2), Adequately (3), and Very Well (4). In addition, there was one Yes/No question and two constructed-response items.

Highlights of responses include the following:

- Four of 14 respondents had used the WAT software before.
- The average responses to the questions about how well the training prepared the group for the various aspects of the alignment process were between 3.00 and 3.50 (out of 4).
- Participants expressed interest in receiving information about the outcomes of the study.

Daily Evaluation of Process Questionnaires

At the end of each day's work, the panelists were asked to complete a survey about how the day had gone, in general. The following table shows the average for each day of the comfort level of the participants with assigning DOK levels and of the perception of how well the facilitator managed the group consensus process.

Table 20: Participants' daily evaluation responses

Survey Item	Monday, 1/25/10	Tuesday, 1/26/10	Wednesday, 1/27/10	Thursday, 1/28/10
1. How comfortable do you feel with the process of assigning DOK levels? **(Scale = 1 – 4)**	2.93	2.92	3.31	3.06
2. How well did your group facilitator facilitate the consensus process? **(Scale = 1 – 3)**	2.89	2.81	2.75	2.56

This summary shows that participants had a fairly high level of comfort or confidence in making judgments about DOK levels (most selected "Comfortable" or "Very Comfortable"). The majority of responses about the consensus process were "Very Well," with some responding "Moderately," indicating that the panelists felt the facilitators managed the discussions well.

Overall, participants noted that it would have been useful to have more time for the various steps of the process, as well as to have had additional examples and practice prior to starting the DOK

and item coding. Panelists also noted that the representation of the NAEP framework/standards was challenging to work with.

Sub-Study Evaluations

Panelists were asked to complete evaluations of each sub-study. The evaluations included the following three constructed-responses questions, one Likert item, and a comments section:

- A. For each standard, did the items cover the most important topics you expected by the standard? If not, what topics were not assessed that should have been?
- B. For each standard, did the items cover the most important performance (DOK levels) you expected by the standard? If not, what performance was not assessed?
- C. Were the standards written at an appropriate level of specificity and directed towards expectations appropriate for the grade level?
- D. What is your general opinion of the alignment between the standards and assessment? (Not at All Aligned; Minimally Aligned; Moderately Aligned; Highly Aligned)
- E. Comments

Evaluation of Sub-Study 1 — NAEP-to-NAEP

Coverage of the standards by the items: Panelists' opinions on how well the items covered the standards varied. Some felt that the items did cover the standards well, while others pointed out a variety of areas of the standards they felt were not covered well. Panelists also discussed the challenge they experienced in identifying appropriate objectives to which to code items.

DOK levels: Many panelists pointed out the lack of DOK Level 4 items. There was also a range of opinion as to whether the DOK levels of the items were appropriate or as expected.

Standards: Again, there was a wide range of opinion expressed, ranging from "Yes, I think the objectives were reasonably specific" to "I would encourage a major reworking of the standards." Others noted that there is a difference between "specific" and "clear" when it comes to test standards and that specificity may not always ensure clarity.

Alignment: Panel 1 results indicate 17% felt there was acceptable alignment and 83% felt there needs to be slight improvement to the alignment. Panel 2 results indicate that 14% felt there was acceptable alignment, 57% felt there needs to be slight improvement, and 29% felt there needs to be major improvement.

Evaluation of Sub-Study 2 —WorkKeys-to-NAEP

Coverage of the standards by the items: In general, the WorkKeys items covered the NAEP objectives to the degree the panelists expected, focusing on main idea, author's purpose, and specific details in informational documents. The panelists noted that there was a significant amount of the NAEP framework not covered by the WorkKeys items.

DOK levels: Panelists pointed out that many WorkKeys items had DOK levels of 1 and 2. Some stated they expected a bit more representation at DOK Level 3 than they saw.

Standards: Panelists pointed out the differences in the intent and audience of the two assessments and had difficulty answering this question.

Alignment: Panel 1 results indicate 60% felt there was acceptable alignment, 20% felt there needs to be slight improvement to the alignment, and 20% felt there needs to be major improvement. Panel 2 results indicate that 14% felt there was acceptable alignment, 71% felt there needs to be slight improvement, and 14% felt there needs to be major improvement. However, as one panelist pointed out, "Given that WorkKeys is not intended to cover all the NAEP standards, I think the alignment is acceptable."

Evaluation of Sub-Study 3 — NAEP-to-WorkKeys

Coverage of the standards by the items: Panelists' opinions were fairly uniform about the fact that the NAEP items were not well aligned with the WorkKeys standards. In the words of one panelist, "The question is phrased in a manner that seems to expect that I would think there is a match, and that is not at all correct. … [T]he objs. of the WorkKeys should not closely align … as the two tests [i.e., the NAEP passage/items and the WorkKeys objs.] are asking different questions and looking for readers to perform different tasks and also have different expectations …."

DOK levels: In general, panelists either noted that the NAEP items were at somewhat higher DOK levels than the WorkKeys objectives as they'd expected, or that it is not possible to make a meaningful statement in response to this question due to the great differences between the two assessments.

Standards: Panelists noted challenges in coding NAEP items to WorkKeys standards given the differences in the organization and focus of the two frameworks.

Alignment: Panel 1 results indicate 100% felt there needs to be major improvement to the alignment. Panel 2 results indicate that 14% felt there needs to be slight improvement, 71% felt there needs to be major improvement, and 14% felt the two are not aligned in any way.

Evaluation of Sub-Study 4 —WorkKeys-to-WorkKeys

Coverage of the standards by the items: Some panelists commented that there were fewer application items than expected among the WorkKeys items, particularly at the lower levels of the standards. Some commented on how they felt the standards should be changed.

DOK levels: Many panelists indicated that the DOK levels of the WorkKeys items were largely as they had expected.

Standards: Most panelists appear to have felt the standards were written appropriately for the purpose, but some commented that there existed a greater degree of overlap among objectives than they felt there should be.

Alignment: Panel 1 results indicate 17% felt there was a high degree of alignment, 33% felt there was acceptable alignment, and 50% felt there needs to be slight improvement to the alignment. Panel 2 results indicate that 14% felt there was acceptable alignment, 43% felt there needs to be slight improvement, and 43% felt there needs to be major improvement.

Final Mapping Debrief — Mapping Both Assessments to the NAEP Framework

This survey consisted of five constructed-response questions, to which sample panelist responses are shown below.

1) What were major differences between the NAEP and WorkKeys assessment in item types, content coverage, and complexity of items **relative to the NAEP framework**?

- "The NAEP assessment items covered more of the content and complexity defined by the NAEP framework. The NAEP assessment items involved much more complex thought/reasoning (in general) across various types of texts (literary and information). The WorkKeys items also ranged in complexity, but fewer items coded at 'higher' DOK levels. Furthermore, WorkKeys items focused on skills and understanding with informational text."
- "The major differences are: the lack of critique/evaluate items in WorkKeys assessment and the lack of WorkKeys items about author's craft. Conversely, WorkKeys had more 'application' items than the NAEP assessment."

2) In your opinion and based on the content analysis completed for the NAEP framework, what similarities and differences are expected in the content knowledge of students who perform well on each assessment, who perform modestly, and who perform poorly?

- "I would not divide the questions in this way. Students who take NAEP are expected to meet cognitive targets on both literary and non-fiction texts. They are expected to have knowledge and skills that will lead to success in comprehending these materials. Students who take WorkKeys are expected to be able to apply, fairly immediately, what they learn from 'practical' texts such as rules, instructions, legal texts, etc. Students who perform modestly or poorly on either test are assumed to be less capable on the expectations for whichever test they have taken."
- "Those who perform well on the NAEP must be able to express themselves in writing and engage in evaluation and critique. Those who perform well on the WorkKeys can follow complex and poorly written procedures. Those who perform modestly on NAEP can make complex inferences from text. Those who perform modestly on WorkKeys can derive literal meaning from text. Those who perform poorly on NAEP can answer explicit questions. Those who perform poorly on WorkKeys cannot except with the very simplest texts."

3) What similarities were identified between the two assessments?

- "Each test contains a range of item difficultly from relatively easy to challenging. Both tests have some selected response items. Both tests are based upon passages that the test taker is to read and understand. Both tests have an underlying structure. Both tests contain some 'practical' texts."

4) What differences were identified between the two assessments?

- "WorkKeys required a much more 'applied' reading than NAEP. Whereas, the NAEP had a high level of expectation regarding a person's ability to move beyond explicit interpretation of various levels of literal and informational text, WorkKeys focused on pulling important information, making some inferences, and applying to some degree this

information with practical informational text. Additionally, I would note that the vocabulary items seemed to assess vocabulary better on WorkKeys than on NAEP. Clarification of intent in some areas is needed on both frameworks."
- "NAEP had more open-ended questions to provide more flexibility. NAEP had longer passages and often more questions assigned to one particular passage. NAEP had a couple of occasions where multiple passages were provided for analysis. WorkKeys had short but sometimes convoluted passages."

5) Please provide any feedback on the usability of the NAEP framework and WorkKeys specifications documents for this alignment task.

- "A major problem lies in the differences in the frameworks between the two tests. The WorkKeys framework is based upon text difficulty, the NAEP framework on cognitive targets. For this reason it is difficult for coders to assign NAEP items to WorkKeys objectives and vice versa. This leads to a large number of uncodable items."
- "1) Many NAEP items could be coded to many objectives making it hard to come to agreement among panelists toward 1 objective. Framework should reflect items better; all items should be developed from the framework. 2) WorkKeys has redundancy in vocab objectives."

Final Mapping Debrief — Mapping Both Assessments to the WorkKeys Framework

This survey consisted of five constructed-response questions, to which sample panelist responses are shown below.

1) What were major differences between the NAEP and WorkKeys assessment in item types, content coverage, and complexity of items **relative to the WorkKeys framework**?

- "WorkKeys framework stresses application, changing conditions in employment situations, whereas NAEP items focus on academic tasks such as locating, integrating, evaluating. WorkKeys assessments involved items that assessed info in the text — not necessarily the application of that info, though."
- "Many NAEP items were uncodable because the WorkKeys framework had no objectives for literary text. Furthermore the WorkKeys framework basic structure is predicated on difficulty of text. In coding NAEP items to the WorkKeys framework it was necessary to first determine the difficulty of the text on which items were based. However, many times the higher level on WorkKeys framework did not have objectives at the level of the item."

2) In your opinion and based on the content analysis completed for the NAEP framework, what similarities and differences are expected in the content knowledge of students who perform well on each assessment, who perform modestly, and who perform poorly?

- "NAEP is measuring the performance from bottom to top of a scale. WorkKeys is trying to measure only what is necessary for satisfactory performance in the workplace eliminating the top of the scale. Someone who does well on WorkKeys will have satisfactory workplace performance. Someone not doing well on WorkKeys is not suitable for workplace. NAEP – top of scale is advanced, bottom is below basic, middle – proficient."

- "Successful performers on the WorkKeys test would be able to apply the content of various passages to projected problems or tasks in the real world. Less successful ones would be less able to make these applications. I would say that students taking NAEP are not so likely to need such in immediate application [of the content of the reading passages to real-world tasks]."
- "I would expect that students who perform well are good readers on informational text — they are able to scan/skim for information. They would also have "good" vocabulary skills. Those who perform poorly would have little exposure to informational text — they would struggle with text at more complex levels (6 – 7). Those who perform modestly would most likely do better on vocab items, but still struggle with complex documents and corresponding items."

3) What similarities were identified between the two assessments?
   - "Literal recall/recognition. Paraphrasing inferences. Main idea/overall purpose. Vocabulary emphasis"
   - "An emphasis on vocabulary in assessments (specifically vocabulary in context). An attempt to move beyond the literal level. Variability in difficulty level of text."

4) What differences were identified between the two assessments?
   - "Purposes and anticipated level of background knowledge. Level of complexity in test items. Different number of test items per passage. Length of texts."
   - "…[T]he focus on informational and procedural documents in WorkKeys — and the representation of literary texts in NAEP (corresponding items referring to author's craft)."

5) Please provide any feedback on the usability of the NAEP framework and WorkKeys specifications documents for this alignment task.
   - "Both worked well, except: The NAEP framework, as it was summarized on the green sheet, was difficult to use as no 1:1 correspondence exists between items and the rows on that chart. The WorkKeys sheet made the distinction among levels less dependent on the difficulty of text than the levels seemed to be. That made coding more difficult."
   - "The frameworks for these tests are not especially useful for this task because neither is meant to be a complete specification of the tests' content. We were finding gaps in how items represented a tests' framework that at times were likely not gaps, but coverage that was meant to be implied."

End-of-Study Questionnaire

Panelists were asked to respond to a final questionnaire upon completion of the entire study. The first seven items were Likert items with four answer options, and the results are shown in Table 21.

Table 21: End-of-study questionnaire summary

Survey Questions	**Average (Scale = 1 – 4)** Averages were calculated by assigning numeric values 1-4 to the response options, in the order shown. Where respondents marked between anchors, a value was assigned and used in the calculations. Responses of "No Answer" are shown, but not included in averages.
1. How well do you feel Monday's training prepared you for understanding depth-of- knowledge (DOK) levels?	3.10
2. How comfortable did you feel with the process of assigning the DOK levels?	3.17
3. How well do you feel Monday's training prepared you for the consensus process?	3.11
4. Overall, how well did Monday's training prepare you for the Alignment (coding) process?	2.94
5. How useful was information about the study you received prior to this week? *"Nothing much there was related to tasks we did - further no other information provided about true nature of task. I have done frameworks, item writing, specialization and analysis but not alignment."*	2.72
6. How useful were the training and coding materials you received this week? *"**Any items marked N.A. are not applicable because of extensive prior experience with the WAT process and alignment. Also, I was a facilitator and not a coder."*	3.00
7. How qualified did you feel your panel was to conduct this type of alignment?	4.00

On average, panelists' responses were in the "adequate" or "comfortable" range for these seven items.

Regarding the rest of the survey, in general, the panelists held the following views:

- The composition of the panel was effective as-is.
- The facilitators were effective adjudicators.
- The alignment criteria were moderately useful.
- The WAT software was fairly easy to use.
- The alignment process was likely able to adequately capture the similarities and differences between the two assessments, but it was difficult to know for certain without seeing the results.
- The participants' perceptions of the similarities and differences between the two assessments were similar to those expressed in the earlier surveys.
- Logistics of the week-long meeting were very suitable.

The survey comments indicate that the panelists' attitudes about their participation in this study were positive overall.

Summary and Conclusions

Key features of the two assessments and their respective item pools used for this study, as delineated by the blueprint analysis and this study, are shown in Table 22.

Table 22: Key features of the NAEP and WorkKeys assessments

Assessment Feature	NAEP Grade 12 *Reading* Assessment	WorkKeys *Reading for Information* Assessment
Item pool	All 131 items of the 2009 NAEP Grade 12 *Reading* item pool were used for this study.	A pool of 60 items drawn from the operational WorkKeys *Reading for Information* item pool of hundreds of items was used for this study.
Types of reading passages	3 of 15 documents used for this study had a workplace context; 1 was consumer oriented. • 30% literary nonfiction, fiction, or poetry • 31% informational expository • 27% argumentative/persuasive • 12% procedural	All 28 WorkKeys documents used for this study had a workplace context. • 32% policy • 35% instructions • 18% legal document • 15% information
Difficulty of reading passages	The difficulty of all reading passages is grade-12 appropriate.	The difficulty of reading passages ranges from grade 6 to postsecondary.
Types of items/Average DOK level	• 58% multiple choice / 1.74 • 32% short constructed response / 2.64 • 10% extended constructed response / 2.92	• 100% multiple choice / 1.54
Standards on which items are based / Average DOK level	1) Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension. / 1.50 2) Integrate/Interpret: Make complex inferences within and across texts. / 2.71 3) Critique/Evaluate: Consider text(s) critically. / 3.10	3) Individuals read short, simple, and clearly stated materials to find out what should be done. / 1.20 4) Individuals read straightforward information that contains a number of details. When following procedures, they must think about changing conditions that affect what should be done. / 1.75 5) Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions. / 1.83 6) Individuals read elaborate procedures, complicated information, and legal regulations, all of which contain difficult words, jargon, and technical terms. Most information is not clearly stated. / 2.43 7) Individuals read very complex information which includes a lot of details and complicated concepts. Unusual jargon and technical terms are used but not defined. Writing often lacks clarity and direction. Individuals must draw conclusions from some parts of the reading and apply them to other parts. / 2.33

Each of the two concurrent, replicate panels convened for this study demonstrated a high degree of interrater agreement. Furthermore, through the processes of both intra-panel and inter-panel adjudication, the two concurrent panels reached a high degree of agreement on their judgments about the alignment of the NAEP Grade 12 Reading assessment and the WorkKeys *Reading for Information* assessment. Thus, it is reasonable to have confidence in the reliability of each panel's ratings.

The data from the two panels shows the following about the Depth of Knowledge (DOK) levels of the two assessments' standards and items:

- The range of DOK levels assigned to the NAEP standards was 1 – 4, and the average DOK level of the NAEP standards was 2.49.
- The range of DOK levels assigned to the WorkKeys standards was 1 – 3, and the average DOK level of the WorkKeys standards was 1.92.
- The range of DOK levels assigned to the NAEP items was 1 – 3, and the average DOK level for all NAEP items was 2.15
- The range of DOK levels assigned to the WorkKeys items was 1 – 2, and the average DOK level for all WorkKeys items was 1.54
- The difference between the average NAEP standard DOK level and the average NAEP item DOK level was 0.34
- The difference between the average WorkKeys standard DOK level and the average WorkKeys item DOK level was 0.38

Many of the key features included in Table 22 have an impact on DOK levels.

Across the four sub-studies, the NAEP and WorkKeys test items were analyzed for their alignment with the three NAEP standards and the five WorkKeys standards according to four alignment criteria. This produced 64 points for which the degree of alignment was evaluated, using labels of Yes (aligned), Weak, and No (not aligned).

For 51 of these 64 points, the two panels agreed on the degree of alignment. For 11 of the 64 points, one panel's assessment was that there was weak alignment while the outcome of the other panel was either Yes or No. And for the two remaining points — Sub-Study 3, Standard 3, Depth of Knowledge and Standard 4, Categorical Consistency, the panels came to opposite conclusions (Yes vs. No).

After the conclusion of the panel meetings, per the study design, the two panel facilitators conferred about the points of disagreement between the two panels. Serving as representatives of their respective panels' discussion and views, the facilitators attempted to adjudicate all differences. Even despite their thorough effort, there remained the few areas of difference between the panels described in the preceding paragraph.

Following is a summary of the outcome of each sub-study:

Sub-Study 1, the alignment of the NAEP items to the NAEP standards:
Sub-Study 1 results showed that the alignment criteria were met for all three standards with respect to three of the four alignment criteria: Categorical Concurrence, Depth-of-Knowledge Consistency, and Range of Knowledge. The criterion that was not fully met was Balance of Representation, indicating that within each standard, the associated test items are clustered around a limited number of objectives.

Four of 131 NAEP test items (3%) were coded to a generic NAEP standard by at least one panelist, indicating that there was a small number of items that some panelists did not feel aligned precisely to any specific objective. There were no items the panelists deemed uncodable. The objectives that were not targeted by any items included several objectives within all three standards.

Sub-Study 2, the alignment of the WorkKeys items to the NAEP standards:
Sub-study 2 results showed some degree of alignment (aligned or weakly aligned) across all four alignment criteria for the first two NAEP standards. There was no alignment, however, to Standard 3, "Critique/Evaluate: Consider text(s) critically." Some factors contributing to this result are summarized in Table 22 and include the differences in the types of items, the types of reading passages, and the standards on which the items are based for the two assessments.

There were no WorkKeys items coded to generic NAEP standards, and there were no items that were deemed uncodable. NAEP objectives to which no WorkKeys items aligned include those specific to literary text and literary devices, and all of the objectives under Standard 3, "Critique/Evaluate: Consider text(s) critically."

Sub-Study 3, the alignment of the NAEP items to the WorkKeys standards:
Sub-Study 3 results showed some degree of alignment (aligned or weakly aligned) across almost all standards and alignment criteria. The Range-of-Knowledge criterion had the fewest aligned results, with 3 of 10 points showing no alignment, and another 3 of the 10 points showing weak alignment. This result indicates that the NAEP items that were codable to WorkKeys standards were clustered around a limited number of objectives, rather than being spread more evenly among all the objectives within the standards.

In each panel, the participants agreed that approximately 17% of the NAEP items were uncodable to WorkKeys standards (see Table 14), particularly if their focus was on a literary device or on the skill of critique/evaluation, as these topics or skills are not included in the WorkKeys standards. Fifty-six percent of the NAEP items were rated as uncodable to WorkKeys standards by at least one panelist. In addition, 16 NAEP items (12%) were coded to a generic WorkKeys standard by at least one rater, and one NAEP item (0.8%) was coded to a generic standard by all panelists, indicating that there were some items that panelists did not feel aligned precisely to any specific objective. The WorkKeys objectives to which no NAEP items aligned are related to applying instructions, understanding jargon, or understanding the general principles behind policies, rules,

and procedures. As with Sub-Study 2, some factors contributing to these results are summarized in Table 22.

Sub-Study 4, the alignment of the WorkKeys items to the WorkKeys standards:
Sub-Study 4 results showed some degree of alignment (aligned or weakly aligned) across almost all standards and alignment criteria. Results for Standard 5 ("Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions.") show more points of weak alignment (1 of 8 points) or no alignment (3 of 8 points) than the other standards. Table 22 includes item pool as a key feature of the assessments for this study, and it is likely a contributing factor in this result. Item sampling to produce the group of WorkKeys items used in this study involved using two intact WorkKeys test forms of 30 items each (for a total of 60 items), rather than a more extensive sample from the entire WorkKeys pool of hundreds of operational items. This decreased the likelihood that the sampled items would completely cover all WorkKeys objectives. As a point of contrast, 131 NAEP items were used in the study.

No items were rated as uncodable by an entire panel. Three of 60 WorkKeys items (5%) were rated as uncodable by one panelist each. In addition, there was one item (1.7%) that was aligned to a generic standard by all panelists, indicating that panelists did not feel this item aligned precisely to any particular objective. Most of the WorkKeys objectives to which one or both panels did not code WorkKeys items are related to vocabulary.

Assessment-to-Assessment Alignment Summary

The following two tables summarize the four sub-studies together. Table 23 shows the distribution of the test content. For each sub-study, the percentage of hits for codable items is shown for each standard. Note that this table shows which standards the test items were coded to; it does not indicate the distribution of items among the objectives within each standard. It also does not include information about items that were judged by the panelists to be uncodable to any of the objectives or standards for the test in question.

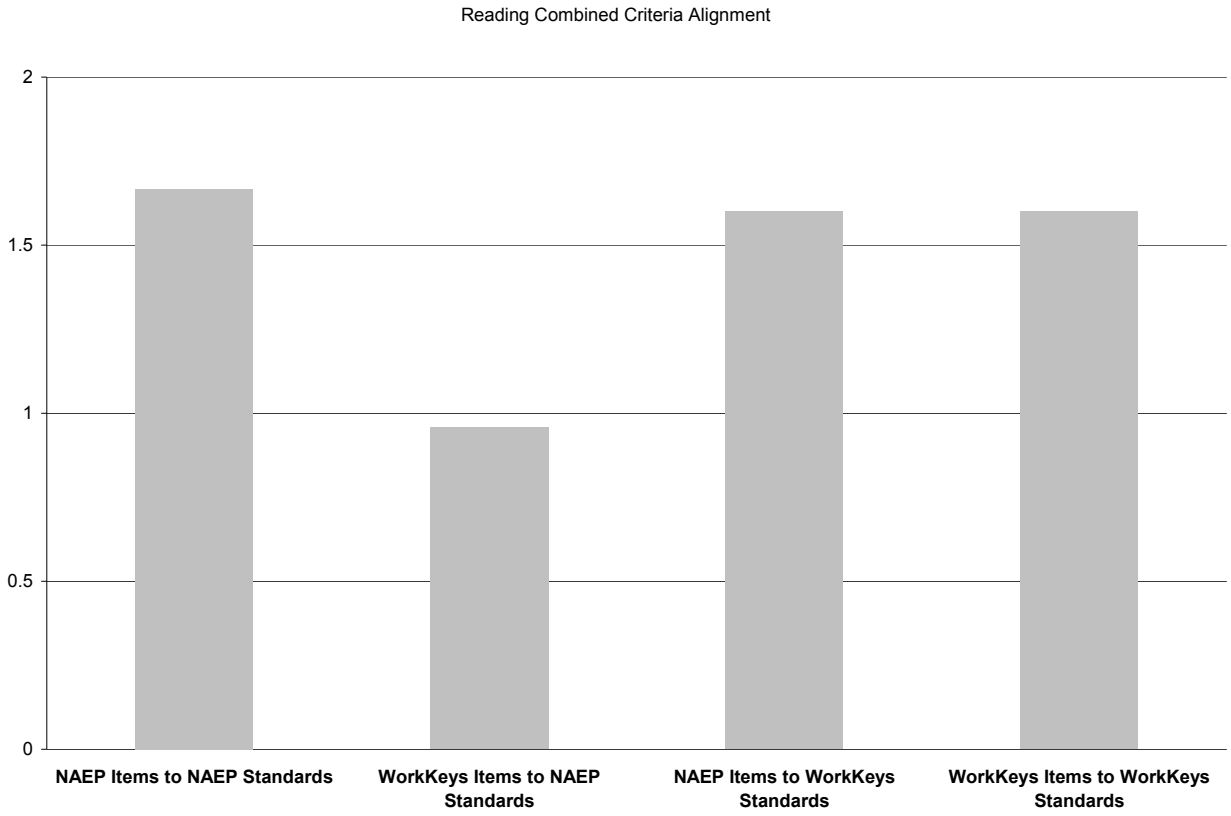
Table 23: Content distribution summary*

NAEP Standards	NAEP Items		WorkKeys Items	
	Panel 1	Panel 2	Panel 1	Panel 2
	Sub-Study 1: % of Hits for Codable Items		Sub-Study 2: % of Hits for Codable Items	
1) Locate/Recall: Locate or recall textually explicit information within and across texts, which may involve making simple inferences as needed for literal comprehension.	28%	28%	70%	71%
2) Integrate/Interpret: Make complex inferences within and across texts.	56%	59%	30%	29%
3) Critique/Evaluate: Consider text(s) critically.	16%	13%	0%	0%
WorkKeys Standards	**NAEP Items**		**WorkKeys Items**	
	Panel 1	Panel 2	Panel 1	Panel 2
	Sub-Study 3: % of Hits for Codable Items		Sub-Study 4: % of Hits for Codable Items	
3) Individuals read short, simple, and clearly stated materials to find out what should be done.	30%	30%	24%	29%
4) Individuals read straightforward information that contains a number of details. When following procedures, they must think about changing conditions that affect what should be done.	6%	7%	20%	20%
5) Individuals read information that is stated clearly and directly, but includes many details, jargon, technical terms, acronyms, or words with several meanings. Individuals typically apply information to a situation not specifically described. They may need to consider several things in order to choose the correct actions.	12%	15%	8%	8%
6) Individuals read elaborate procedures, complicated information, and legal regulations, all of which contain difficult words, jargon, and technical terms. Most information is not clearly stated.	38%	37%	43%	39%
7) Individuals read very complex information which includes a lot of details and complicated concepts. Unusual jargon and technical terms are used but not defined. Writing often lacks clarity and direction. Individuals must draw conclusions from some parts of the reading and apply them to other parts.	15%	11%	5%	5%

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

The next graphic, Table 24 compares all four sub-studies when the four alignment criteria are combined and considered together. The graph was calculated by assigning a point value of 2 to each analysis point that was aligned ("Yes"), 1 point to each analysis point that was weakly aligned ("Weak"), and 0 points to each analysis point that was not aligned ("No") The sum was then divided by the total possible points for the study (24 points [12 analysis points X 2] for Sub-Studies 1 and 2, and 40 points [20 analysis points X 2] for Sub-Studies 3 and 4.

Table 24: Combined alignment criteria



The values shown in Table 25 were used to create the graph in Table 24.

Table 25: Combined alignment criteria data

Sub-Study	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Represent-ation	Combined
1 (NAEP items to NAEP standards)	2.00	2.00	2.00	0.67	1.67
2 (WorkKeys items to NAEP standards	1.33	1.17	0.33	1.00	0.96
3 (NAEP items to WorkKeys standards)	1.80	1.80	1.10	1.70	1.60
4 (WorkKeys items to WorkKeys standards	1.20	1.70	1.70	1.80	1.60

General conclusions

The following conclusions are supported by the data from these studies:

A) The NAEP Grade 12 Reading assessment covers a broader range of reading skills than does the WorkKeys *Reading for Information* assessment, particularly in the literary genre and in requiring examinees to critique and evaluate reading materials.

B) The WorkKeys *Reading for Information* assessment focuses on a narrower range of reading skills than does the NAEP assessment. Specifically, the WorkKeys assessment focuses on workplace communications, especially policies and instructions, and their application to workplace situations.

C) Most of the WorkKeys items aligned with NAEP objectives were related to locating/recalling information and causal relations. WorkKeys items that aligned under Standard 2, Integrate/Interpret, targeted objectives that require the examinee to connect ideas, draw conclusions and provide supporting information, and to determine word meaning in context. No WorkKeys items included in this study require the examinee to critique or evaluate the reading passage.

D) WorkKeys objectives that are not assessed by the NAEP items include applying complex, multistep, conditional instructions to similar and new workplace situations; determining the meaning of work-related acronyms, jargon, and technical terms; and figuring out and applying general principles contained in informational documents to similar and new workplace situations.

Contractor Comments on Study Design

The amount of work required of the panelists for this type of alignment-to-alignment study is enormous, and participants must demonstrate noteworthy fortitude to complete the study in a week's time. For this reason, any future revisions of the overall process must not result in more time being required of the panelists, as this would reduce the likelihood of successfully recruiting panelists able to commit more than a week of time and would increase the demands on and, therefore, the fatigue of the participants — which may be counterproductive.

There are two areas of the process ACT particularly recommends reviewing for future applications of this assessment-to-assessment alignment approach. One is the Balance of Representation criterion. The description of this criterion in Dr. Webb's paper describing the alignment methodology appears to be somewhat at odds with the results of the calculations programmed in the WAT software, and the statistical results in this category may be misleading in situations in which the overwhelming majority of panelists code a large number of items to only one objective within a standard. When this aspect of the WAT Balance of Representation calculation is taken into account, the count of aligned, weakly aligned, and not aligned points in the standards becomes less clear.

The other area of the process ACT recommends reviewing for future applications of the approach has to do with just how to make the judgment about the overall alignment of two assessments. In

other words, when aligning a single assessment to a set of standards, the methodology and WAT software include clearly prescribed threshold values for determining the degree of alignment. When two assessments are compared, however, this task becomes much more complex. It may be useful to further explore this type of study to see whether it would be appropriate to establish threshold values or other markers for assessment-to-assessment alignment studies. Again, however, a challenge here would be that whatever might be established for future procedures must not lead to substantial increase in the amount of work or time required of the panelists, as the existing procedures push the outer limits of what is feasible.

Dr. Webb's alignment methodology has been applied to dozens of test-to-standards alignment studies to date, with reputable results. The adaptation of the methodology for an assessment-to-assessment alignment study is new, and this adaptation has also resulted in a great deal of useful and informative data. The WAT software tool, in particular, is instrumental in facilitating the work, allowing study participants to manage a great deal of information and judgment easily and allowing study facilitators to manage the data much more easily and quickly than could be done manually. Even if no changes are made to the methodology in the future, it is clear that a great deal of useful information is created by the process as it currently stands — and that the results of this rigorous work can be mined to reasonably and confidently inform decisions related to the future directions of the testing programs under scrutiny.

References

ACT, Inc. (January, 2010). *Grade 12 NAEP reading assessment and WorkKeys Reading for Information assessment blueprint analysis*. Iowa City, IA.

ACT, Inc. (2008). *Reading for information technical manual.* Iowa City, IA.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Messick, S. (1994, March). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62(3), 229-258.

National Assessment Governing Board. (2008). *Reading framework for the 2009 national assessment of educational progress.* Washington, D.C.: U.S. Department of Education.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25* (1), 47-55.

Valencia, S. W., & Wixson, K. K. (2000). Policy-oriented research on literacy standards and assessment. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (pp. 909-935). Mahwah, NJ: Erlbaum.

Webb, N. L. (2005) *Web alignment tool (WAT) training manual*. Madison, WS: University of Wisconsin, Wisconsin Center for Educational Research.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002). Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states. A study of the State Collaborative on Assessment & Student Standards (SCASS) Technical Issues in Large-Scale Assessment (TILSA). Washington, D. C.: Council of Chief State School Officers.

Webb, N. L. (2009). *Design of content alignment studies in mathematics and reading for 12th grade NAEP preparedness research studies*. Washington, D. C.: U.S. Department of Education.

Appendices

Appendix A: Design of Content Alignment Studies in Mathematics and Reading for 12th Grade NAEP Preparedness Research Studies
Appendix B: Grade 12 NAEP *Reading* Assessment and WorkKeys *Reading for Information* Assessment Blueprint Analysis
Appendix C: Day-by-Day Agenda
Appendix D: Reading Depth-of-Knowledge Training Materials
Appendix E: Inter-Panel Consensus Depth-of-Knowledge Values for the Test Standards
Appendix F: Evaluation Forms
Appendix G: Panelists' Responses to Evaluation Forms
Appendix H: NAEP Item DOK Levels
Appendix I: WorkKeys Item DOK Levels
Appendix J: WAT Reports — Sub-Study 1, Panel 1
Appendix K: WAT Reports — Sub-Study 1, Panel 2
Appendix L: WAT Reports — Sub-Study 2, Panel 1
Appendix M: WAT Reports — Sub-Study 2, Panel 2
Appendix N: WAT Reports — Sub-Study 3, Panel 1
Appendix O: WAT Reports — Sub-Study 3, Panel 2
Appendix P: WAT Reports — Sub-Study 4, Panel 1
Appendix Q: WAT Reports — Sub-Study 4, Panel 2
Appendix R: WAT Reports — Cross-Study Table for NAEP Objectives
Appendix S: WAT Reports — Cross-Study Table for WorkKeys Objectives
Appendix T: Study Participants and ACT Project Staff
Appendix U: Explanation of Rater Agreement Statistics