The Alignment of the NAEP Grade 12 *Mathematics* Assessment and the WorkKeys *Applied Mathematics* Assessment

August 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
Linda McQuillen
Linda Wilson

Study panelists
[redacted]

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 *Mathematics* assessment and ACT, Inc.'s WorkKeys *Applied Mathematics* 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 primarily over the course of a week in January 2010 at ACT's national headquarters in Iowa City, IA, using two concurrent, replicate panels of math content experts from across the United States. A final portion of the study was conducted remotely in February 2010.

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, straightforward recall and skills to deep knowledge and higher-order thinking skills. Mathematics assessment materials at DOK Level 1 typically involve a rote response, performing a simple algorithm, or applying a simple formula. Those at DOK Level 2 involve both comprehension and subsequent mental processing requiring more than one step, such as classifying, estimating, or comparing. At DOK Level 3, mathematics assessment materials typically require reasoning, planning, using evidence, and making conjectures. DOK Level 4 mathematics assessment materials require complex reasoning, planning, developing, and thinking, most likely over an extended period of time; they may involve designing and conducting experiments and critiquing experimental designs. 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 – 3, and the average DOK level of the NAEP standards was 2.00.
- The range of DOK levels assigned to the WorkKeys standards was 1 – 3, and the average DOK level of the WorkKeys standards was 1.58.
- The range of DOK levels assigned to the NAEP items was 1 – 3, and the average DOK level for all NAEP items was 1.90.
- The range of DOK levels assigned to the WorkKeys items was 1 – 2, and the average DOK level for all WorkKeys items was 1.97.

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 *Mathematics* Assessment	WorkKeys *Applied Mathematics* Assessment
Item pool	All 178 items of the 2009 NAEP Grade 12 *Mathematics* item pool were used for this study.	A pool of 58 items drawn from the operational WorkKeys *Applied Mathematics* item pool of hundreds of items was used for this study.
Item context	Items include both pure math and real-world content	All items involve real-world application of math content in a workplace context
Types of items/Average DOK level	• 61% multiple choice / 1.85 • 29% constructed response / 2.14 • 11% multiple part / 1.95 NOTE: Percentages do not equal 100% due to rounding.	• 100% multiple choice / 1.97
Standards on which items are based / Average DOK level	1) Number properties and operations / 1.80 2) Measurement / 1.94 3) Geometry / 2.00 4) Data analysis, statistics, and probability / 2.16 5) Algebra / 2.00	3) A single type of basic mathematics operation; no reordering or extraneous information / 1.20 4) Multiple types of mathematical operations; reordering, extraneous information / 1.29 5) Application of logic and calculation; conversions / 1.29 6) Complex and multiple-step calculations; manipulating formulas / 1.78 7) Nonlinear functions, complex calculations and conversions / 2.14

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 five NAEP standards and the five WorkKeys standards according to four alignment criteria. This produced 80 points for which the degree of alignment was evaluated, using labels of Yes (alignment), Weak, and No (not aligned); the results of each sub-study are given in detail in the body of this report. The two concurrent panels reached the same conclusions for 78 of these points, showing a high degree of consistency between the two panels. For one of the remaining two points, the panels reached a similar conclusion (no alignment vs. weak alignment, with similar index values), and for the other point, they reached opposite conclusions (no, not aligned, vs. yes, aligned). However, this instance of opposite conclusions is the result of the judgment of one panelist regarding one test item and is explained in more detail in the discussion of Sub-Study 2 in the body of this report.

In general, study results, including the blueprint analysis, showed that the NAEP assessment covers a broad range of content that represents the typical high school mathematics curriculum. In contrast, the WorkKeys assessment covers a narrower range of content that focuses on applying foundational math skills in workplace situations. In one sense, the content represented by the WorkKeys standards and items may be considered a subset of the content represented by the NAEP standards. And, in fact, study results showed that all of the WorkKeys test items used for this study may be coded to a subset of NAEP objectives. However, the blueprint analysis and sub-study results also showed that the NAEP standards and items used for this study do not cover all of the objectives within the WorkKeys standards, indicating that there is content represented by the WorkKeys standards that is not covered by the NAEP assessment.

The NAEP Grade 12 *Mathematics* framework is organized into five content-oriented standards: Number Properties and Operations; Measurement; Geometry: Data Analysis, Statistics, and Probability; and Algebra. Among these five standards are 154 objectives. This study used the full 2009 item pool of 178 NAEP items.

Sub-Study 1 found alignment of the NAEP items to the NAEP standards at all points for which alignment was calculated. Sub-Study 3 found alignment of the NAEP items to the WorkKeys standards at 75% of the points for which alignment was calculated. The WorkKeys objectives that were most frequently covered by NAEP items used for this study include geometry content; fractions, ratios, percentages, or mixed numbers; and basic statistical concepts. The WorkKeys objectives that were targeted least often or not at all by the NAEP items range across a variety of workplace-oriented math skills: conversions, determining the best deal, finding errors, and calculating discounts or markups.

The WorkKeys *Applied Mathematics* test framework is organized into five levels for which skills and the characteristics of items are defined and contextualized for the workplace. The five levels are described as follows: 3 — A single type of basic mathematics operation; no reordering or extraneous information; 4 — Multiple types of mathematical operations; reordering, extraneous information; 5 — Application of logic and calculation; conversions; 6 — Complex and multiple-step calculations; manipulating formulas; 7 — Nonlinear functions, complex calculations and conversions. Among these five standards are 34 objectives. This study used the items from two intact test forms. Each form has 30 operational items, and the two forms used in this study had two items in common, for a total of 58 unique items out of the full item pool of hundreds of operational WorkKeys items.

Sub-Study 4 found alignment of the WorkKeys items to the WorkKeys standards at 85% of the points for which alignment was calculated. Sub-Study 2 found alignment of the WorkKeys items to the NAEP standards at 40% of the points for which alignment was calculated. The NAEP objectives targeted by the most WorkKeys items included problem-solving applications of number operations and measurement; those to which no WorkKeys items aligned were primarily related to geometry, data analysis, statistics, probability, and algebra.

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 overlap between the content represented by the two tests, there are also important differences. In addition to the differences in content that were revealed by the study, the panelists commented on the differences in the purposes of the two assessments. The NAEP assessment is intended to measure twelfth-grade student achievement in mathematics, while the WorkKeys assessment is intended to measure skill in applying mathematical reasoning and problem-solving techniques to work-related problems.

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 Mathematics assessments as administered in 2009 and the content of the WorkKeys *Applied Mathematics* 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 mathematics assessments, as well as the cognitive complexity of these assessments.

Methodology

Study Design

The Webb alignment methodology used for this study specifies that an independent content expert should conduct an analysis of the test blueprints prior to assembling the content experts for the alignment study. Accordingly, an expert in math first analyzed the NAEP and WorkKeys test blueprints to identify similarities and differences in the respective tests' specifications. This analysis found that there is overlap in the content assessed by both tests, but the NAEP assessment covers a broader range of cognitive targets than the WorkKeys assessment does. Both tests target number properties and operations, measurement, and geometry. The other content areas included on the NAEP assessment are targeted either very little or not at all by the WorkKeys assessment; these areas include data analysis, statistics, probability, and algebra. 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 *Mathematics* assessment and the WorkKeys *Applied Mathematics* assessment.

Table 1: Comparison of the critical features of the NAEP Grade 12 Mathematics test and the WorkKeys Applied Mathematics test, excerpted from blueprint analysis report

	NAEP GRADE 12 *MATHEMATICS* ASSESSMENT	WORKKEYS *APPLIED MATHEMATICS* ASSESSMENT
Number of Items	• Total number of items is 150 to 200; a matrix sample design is used so that each examinee receives 40 to 45 items	• 33 items, 30 of which are scored and three of which are unscored pretest items
Item Formats	• Test includes multiple-choice items (5 choices) and both short and extended constructed-response items • Test time (not number of items) is divided equally between multiple-choice and constructed-response items	• 100% multiple choice (5 choices)
Item Content	• Items include both pure math and "real-world" content • Content is divided up as follows: – Number properties and operations: 10% – Measurements with geometry: 30% – Data analysis/ statistics/ probability: 25% – Algebra: 35%	• All items involve real-world application of math content in a workplace context • Content is divided up as follows: – Number properties and operations: 62% – Measurements with geometry: 32% – Data analysis/ statistics/ probability: 0 – 6% * – Algebra: 0 – 6 % * * Found in some higher-level items only
Item Complexity	• 3 levels: – Low (25%) — recall or recognize concepts or procedures; carry out specified procedures – Moderate (50%) — think flexibly in solving problems; make connections among concepts and processes from various domains – High (25%) — use reasoning, planning, analysis, judgment, and creative thought in solving problems	• 5 levels: – Level 3 – basic mathematics operations (20%) – Level 4 – multiple mathematical operations (20%) – Level 5 – application of logic and calculation (20%) – Level 6 – complex and multiple-step calculation (20%) – Level 7 – nonlinear functions, conversions, complex calculations (20%)

	NAEP GRADE 12 *MATHEMATICS* ASSESSMENT	WORKKEYS *APPLIED MATHEMATICS* ASSESSMENT
Resources	- Calculator is permitted for 1/3 of the total item blocks and is not permitted for 2/3 of the blocks (no QWERTY keyboards); items for blocks that permit calculator use are rated on the degree to which having a calculator is useful to the student solving the problem; no items are designed to provide an advantage to students using a graphing calculator over students who are not. - Manipulatives and mathematical tools are available for selected items and include the following: – number tiles – geometric shapes – rulers – protractors	- Available for entire test: – Calculator (no QWERTY keyboards) – Formula sheet
Assessment Time	- 50 minutes, divided into two blocks of 25 minutes each	- 55 minutes when computer-delivered - 45 minutes when given by paper and pencil
When Given	- Every four 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). Beginning in 2009 on a voluntary basis, 12th-grade testing was conducted at the state level. - The samples of students are designed to be representative of the nation and are drawn from different regions of the country and participating states. - Before the 2011 administration, ELL students participated unless they had less than 3 school years of instruction in English. Beginning in 2011, ELL students will participate unless they have had less than 1 school year of instruction in English.	- High school students, job applicants, current employees, people seeking certification or other documentation of their skill levels. Approximately 780,000 WorkKeys *Applied Mathematics* tests were administered in fiscal year 2009.

	NAEP GRADE 12 *MATHEMATICS* ASSESSMENT	WORKKEYS *APPLIED MATHEMATICS* ASSESSMENT
Accommodations	• Allow accommodations specified in an IEP that are routinely used in testing, including but not limited to: – one-on-one testing – small-group testing – extended time – oral reading of directions – large-print booklets – bilingual English/Spanish booklets – use of an aide to transcribe responses • For a complete list of NAEP math 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: 0 to 2 points (either correct/incorrect; or correct, partially correct, incorrect) • Extended constructed response: 0 to 4 points – 4 — extended – 3 — satisfactory – 2 — partial – 1 — minimal – 0 — incorrect	• Multiple choice: – Incorrect 0 – Correct 1 • No penalty for guessing

	NAEP GRADE 12 *MATHEMATICS* ASSESSMENT	WORKKEYS *APPLIED MATHEMATICS* ASSESSMENT
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 knowledge and skill fundamental for proficient work at grade level • **Proficient** — Solid academic performance for grade; competency over challenging subject matter and application to real-world situations and analytical skill appropriate for subject matter • **Advanced** — Superior performance Test scores are not determined for individual examinees. 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:** Score range is Levels 3 through 7; a score of Below 3 also may be given. Level 3 is the lowest level generally useful in a job; individuals possessing math skills below this level are generally not qualified for jobs that require even the most basic math, and employers are typically not willing to train individuals with math 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 11 – 15, 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 this 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. The pilot used the reading content rather than math content. However, the same alignment methodology and meeting procedures were to be used for both content areas, so it was believed that lessons learned from the reading pilot generally would be applicable to the math study, as well.

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 reading 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 — After the pilot, in preparation for the math panel in January, five WorkKeys math test items were added to the DOK training packet because the sample items provided in the training materials were more similar to the NAEP items than to the WorkKeys items. Feedback from the pilot indicated that it would help the panelists assign DOK levels more accurately to the WorkKeys items if there were more WorkKeys-specific practice provided.

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 "objective" 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?

For these questions, we concluded that the participants in the full study 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 mathematics 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. However, due to scheduling, the study included only one panelist who was a member of a minority racial or ethnic group. 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, the two panels included seven and eight 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 the *Mathematics Framework for the 2009 National Assessment of Educational Progress* (September, 2008), Exhibits 3 – 7. Each exhibit describes a content area representing a standard. Within these exhibits are subtopics representing the goals for the standard. Under the subtopic is a table with three columns representing Grade 4, Grade 8, and Grade 12. The cells contain the objectives. Because this study is focused on Grade 12, only the objectives for Grade 12 were used. In order to fit within the constraints of the WAT, the exhibits from the NAEP Framework document was translated into an outline format.

Following is an excerpt from the NAEP Grade 12 *Mathematics* assessment standards used in this study. Throughout this report, the text describing the content area 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 subtopic of the content area 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.d) and is referred to as the *objective*. The lettering used in the NAEP Framework document was retained for this study, including some gaps in alphabetical lettering, as shown in Table 2.

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

Level	Description
1	Number Properties and Operations
1.1	Number sense
1.1.d	Represent, interpret, or compare expressions for real numbers, including expressions using exponents and logarithms.
1.1.f	Represent or interpret expressions involving very large or very small numbers in scientific notation.
1.1.g	Represent, interpret, or compare expressions or problem situations involving absolute values.
1.1.i	Order or compare real numbers, including very large and very small real numbers.
1.2	Estimation
1.2.b	Identify situations where estimation is appropriate, determine the needed degree of accuracy, and analyze the effect of the estimation method on the accuracy of results.

The WorkKeys *Applied Mathematics* 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 math items vary in complexity according to the skill level. The WorkKeys Applied Mathematics 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 Applied Mathematics 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 *Applied Mathematics* assessment and the WorkKeys program were consulted to ensure that the adapted standards were consistent with the original skill definitions.

Following is an excerpt from the WorkKeys *Applied Mathematics* 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* Applied Mathematics *standards

Level	Description
3	A single type of basic mathematics operation; no reordering or extraneous information
3.1	Solve problems that require a single type of mathematics operation (addition, subtraction, multiplication, and division) using whole numbers
3.2	Add or subtract negative numbers
3.3	Change numbers from one form to another using whole numbers, fractions, decimals, or percentages
3.4	Convert simple money and time units (e.g., hours to minutes)
4	Multiple types of mathematical operations; reordering, extraneous information
4.1	Solve problems that require one or two operations
4.2	Multiply negative numbers

Assessments

The NAEP Grade 12 *Mathematics* 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 178 items for the 2009 assessment provided by NAGB for this study.

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 as an assessment 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 *Applied* Mathematics 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 *Applied Mathematics* forms consisting of 30 operational items each. Items on the second form were renumbered so that they would be in sequence following the first form. All items are set in a workplace context. All 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. Because of the way that WorkKeys assessments are equated, the two test forms used for this study included two items in common. These items were included only once each for the study, so the total number of WorkKeys items in the study was 58.

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 *Applied Mathematics* 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 *Applied Mathematics* 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 mathematics 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 mathematics. 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 mathematics 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 log-ins, 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 Mathematics *items to NAEP Grade 12* Mathematics *standards*
Assign DOK levels to each objective in the NAEP framework
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 framework
Adjudicate mapping within each panel
Complete evaluation of just-completed work

Sub-Study 2: WorkKeys Applied Mathematics *items to NAEP Grade 12* Mathematics *standards*
Assign DOK levels to WorkKeys items
Map WorkKeys items to NAEP framework
Adjudicate mapping within each panel
Complete evaluation of just-completed work

Sub-Study 3: NAEP Grade 12 Mathematics *items to WorkKeys* Applied Mathematics *standards*
Assign DOK levels to each objective in the WorkKeys framework
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 framework
Adjudicate mapping within each panel
Complete evaluation of just-completed work

Sub-Study 4: WorkKeys Applied Mathematics *items to WorkKeys* Applied Mathematics *standards*
Map WorkKeys items to WorkKeys framework
Adjudicate mapping within each panel
Complete evaluation of just-completed work
NOTE: This study was completed remotely in February as described in the next section, "Decision Rules and Adjudication," Number 8. The reason for this was that the very large number of items and standards to be analyzed was simply too much for the panelists to complete in the week allotted.

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

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) Activity was done in a different order than stated on the agenda — the groups coded WorkKeys items to NAEP standards first, rather than coding NAEP items first; and not all activities were finalized before panelists moved on to the next activity.

Rationale: Throughout the entire week's work, the amount of time available to complete the required tasks was a primary, driving concern. After coding and adjudicating DOK levels for the NAEP standards, the facilitators recommended coding the WorkKeys items to NAEP standards first, rather than coding the NAEP items first as originally planned. The facilitators believed that the smaller number of WorkKeys items (58, in contrast with the 178 NAEP items) would be a more manageable way for the panelists to learn and gain speed in the item-coding process.

In addition, the facilitators looked for ways to get the panelists started working on the next task, while they (the facilitators) reviewed data and identified areas that required further discussion and adjudication. Once the facilitators had completed their review and planning, they called their groups back to the previous task for the necessary discussion. Once the task was finalized, the groups returned to their work on the current activity. The result was that the panelists had very few breaks in order to improve the group's chances of completing all the work required.

2) Fifty-eight WorkKeys items were used for this study rather than 60, as originally anticipated.

Rationale: ACT originally communicated to NAGB that there would be two intact WorkKeys test forms used for this study, with a total of 60 operational test items. However, due to WorkKeys test development methodology, in which common items are used to equate different test forms, the two test forms selected for the study contained two common items. ACT chose not to replace these items with other, unique items in the interest of using authentic, intact test forms for this study. Therefore, the total number of unique WorkKeys items used was 58 instead of 60.

3) The two panel groups were brought together to discuss and adjudicate DOK levels of NAEP standards after achieving consensus within individual groups, rather than solely having the facilitators make this decision.

Rationale: The Webb methodology specifies that the two replicate panels will work independently to come to consensus within themselves on the DOK levels of the test standards. After that, if there are any differences between the two panels in the DOKs assigned, the facilitators are to make the judgment as to which level will be assigned. The end result is to be that the two panels use the same DOK levels for the standards.

On the first day of the meeting, instead of having the facilitators do all of this adjudicating, the two panels were brought together to discuss the differences in the two groups' assigned DOK levels for the NAEP standards and to reach consensus among the full group (all 15 panelists).

This was done at the recommendation of the facilitators as a training exercise, a sort of investment in the rest of the week. The facilitators felt that it would be beneficial for the panelists to participate in the discussion and reasoning about which DOK levels were appropriate for the objectives in question. Their experience was that the process of internalizing the DOK levels and being able to consistently apply them accurately takes time and guidance.

The combined group did not finish adjudicating the DOK levels for all of the objectives in question, so the facilitators adjudicated the remainder that evening, after the group had adjourned for the day. The panelists' assessment of this experience was that it was very helpful in furthering their DOK understanding and accuracy. The following day's results also suggested that the exercise had been helpful, as the agreement among panelists improved. Further, this approach was used again in adjudicating the DOK levels for the WorkKeys standards and served as a re-calibration exercise and an opportunity to ensure that all panelists were interpreting the standards in the same way.

4) Panelists were provided information about discrepancies in the NAEP item-to-NAEP standards coding in print, rather than only orally.

Rationale: The very high number of test items and objectives, along with the type of thinking required to code the items meant that the panelists were very pressed for time to complete the required tasks. In addition, the very high number of NAEP objectives resulted in a large number of items for which the differences among the panelists' coding judgments fell outside the parameters specified for the study. One group had 76 items to adjudicate, and the other had 95 — a time-consuming proposition.

Going through the entire adjudication process orally requires more time than does allowing panelists to review the information on paper. Giving the panelists the printout from the WAT of all the items requiring adjudication gave them the same information they would have received orally from the facilitator, but it permitted them to review the items more quickly, to make preliminary decisions about whether they felt their responses should be changed, and to prepare for a more focused group discussion. This ultimately reduced the amount of time required for adjudication of NAEP items-to-NAEP standards discrepancies.

5) The evaluations completed after the panelists finished coding an assessment's items to a set of standards were completed electronically in the WAT, rather than by paper as originally planned.

Rationale: The Webb methodology specifies that evaluation be completed at twelve points during the alignment study, in order to get a sense of how panelists are perceiving the process, whether there are particular problems to address, and what the panelists' perceptions of the nature of the alignment between the two assessments. ACT agreed to use the same evaluation forms as those used by another contractor for the NAGB research studies, WestEd, for similar studies. During the ACT study, however, we decided not to use the paper forms for four of the twelve evaluations, because the survey questions are part of the WAT software. We opted to retrieve the information from the WAT rather than having the panelists write down their answers to the questions on paper.

6) Decision rule for coding DOK levels for WorkKeys standards: Consider each line item independently, not in conjunction with the main level descriptor.

Example: Objective 6.2, “Rearrange a formula before solving a problem,” should be considered and coded independently, not in the context of or in conjunction with the Standard (level descriptor) 6, “Complex and multiple-step calculations; manipulating formulas.”

Rationale: Panelists discussed and agreed to consider each line item (each objective) independently in the WorkKeys standards when assigning DOK levels. The level descriptors, or standards, are often at a higher DOK than the objectives, and the consensus was that considering each objective independently would lead to more precision in the alignment.

7) 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, Webb’s methodology is not commonly used to align two assessments, and the particular differences between the NAEP and WorkKeys standards are such that panelists and facilitators agreed in discussion that many NAEP items would not be codable to the WorkKeys objectives.

The NAEP standards are organized by content area, such as number properties and operations, geometry, and algebra. In contrast, the WorkKeys standards are organized in a hierarchy of cognitive skills, in which the types of problems to be solved move from simple and straightforward to complex, while encompassing a narrower range of content than that encompassed by the NAEP standards. The WorkKeys test is designed to measure examinees’ problem-solving ability involving foundational math skills in a workplace context, whereas the NAEP test is designed to measure what students have learned over the course of at least three years of the high school math courses of algebra 1 and 2 and geometry, and possibly also more advanced math classes.

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. This would mean, for example, that a NAEP item involving a nonlinear function likely would not be matched to a WorkKeys objective, but it might instead be coded to the standard that reads, “Nonlinear functions, complex calculations and conversions.” However, the panelists’ familiarity with the NAEP and WorkKeys tests led them to conclude that such coding practice often would be misleading, indicating alignment that was not actually present. For instance, a NAEP item requiring examinees to analyze an aspect of a nonlinear function might at first seem to fit the WorkKeys level description standard of “Nonlinear functions, complex calculations and conversions.” However, the WorkKeys objectives falling under this standard clarify that this type of analysis problem is not included on the WorkKeys test. The WorkKeys test does include items that require examinees to work with nonlinear functions, but these are problems that can generally be solved using provided formulas rather than this type of analysis. The two tests ask examinees to perform different types of tasks with nonlinear functions.

The WAT is designed to require 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 type in a note each time would be very time consuming over the course of 178 items, the facilitators provided a template note for panelists to copy in 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."

8) The WorkKeys items-to-WorkKeys standards study was not done on-site during the week of January 11 – 15 as originally planned; instead, it was completed remotely in February.

Rationale: Despite the superb work of the facilitators and the diligent work of the panelists — including working through breaks and starting at 7:00 a.m. Friday morning — there was more work than there was time in which to complete it. This was largely due to the high number of NAEP standards and items. The study team (facilitators and ACT staff) determined that it was advisable to fully complete the first three studies (NAEP items-to-NAEP standards; WorkKeys items-to-NAEP standards; and NAEP items-to-WorkKeys standards) and ensure solid data for all three of those studies than it was to hurriedly complete these studies and partially complete the fourth (WK items-to-WK standards).

The final study was completed remotely in February, remaining as close to the Webb methodology as possible. Due to schedule conflicts, not all panelists were able to participate. Each panel had four of the original panelists and the same facilitator as during the on-site portion of the study.

The panelists completed the coding independently and entered their judgments into the WAT software from their remote locations. Once the panelists had completed their work in the WAT, the facilitators reviewed the data in the WAT and, as they did during the on-site portion of the study, they pointed out areas they felt individual panelists may want to review prior to adjudication. After these steps, the facilitators led the adjudication process via conference call.

A gap in the alignment methodology that resulted from conducting this final sub-study remotely was that only one panelist returned the completed evaluation form for the final debrief for mapping assessments to the WorkKeys framework.

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 or fewer, the WAT indicates "No," the Categorical Concurrence criterion has not been met for that standard. There is not a "Weak" alignment result for this criterion. (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 levels in this section is taken from the materials developed by Dr. Webb:

Level 1 (Recall) includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels, depending on what is to be described and explained.

Level 2 (Skill/Concept) includes the engagement of some mental processing beyond an habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of objects or phenomena and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret," could be classified at different levels depending on the object of the action. For example, interpreting information from a simple graph, or reading information from the graph, also are at Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Level 2 activities are not limited only to number skills, but may involve visualization skills and probability skills. Other Level 2 activities include noticing or describing non-trivial patterns, explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and deciding which concepts to apply in order to solve a complex problem.

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

(Depth-of-Knowledge Levels section taken from Webb, 2005, pp. 45 – 46)

Results

Rater Data

Table 4 shows rater agreement statistics 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.

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:* 0.93 *Pairwise Comparison:* 0.79 *Objective Pairwise Comparison:* 0.77 *Standard Pairwise Comparison:* 0.92	*Intraclass Correlation:* 0.92 *Pairwise Comparison:* 0.77 *Objective Pairwise Comparison:* 0.77 *Standard Pairwise Comparison:* 0.92
Sub-Study 2: WorkKeys to NAEP	*Intraclass Correlation:* 0.87 *Pairwise Comparison:* 0.94 *Objective Pairwise Comparison:* 0.77 *Standard Pairwise Comparison:* 0.91	*Intraclass Correlation:* 0.792 *Pairwise Comparison:* 0.94 *Objective Pairwise Comparison:* 0.78 *Standard Pairwise Comparison:* 0.88
Sub-Study 3: NAEP to WorkKeys	*Intraclass Correlation:* 0.93 *Pairwise Comparison:* 0.78 *Objective Pairwise Comparison:* 0.91 *Standard Pairwise Comparison:* 0.93	*Intraclass Correlation:* 0.93 *Pairwise Comparison:* 0.78 *Objective Pairwise Comparison:* 0.88 *Standard Pairwise Comparison:* 0.90
Sub-Study 4: WorkKeys to WorkKeys	*Intraclass Correlation:* 0.76 *Pairwise Comparison:* 0.94 *Objective Pairwise Comparison:* 0.97 *Standard Pairwise Comparison:* 0.98	*Intraclass Correlation:* 0.56 *Pairwise Comparison:* 0.94 *Objective Pairwise Comparison:* 0.97 *Standard Pairwise Comparison:* 0.97

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 4, Panel 2's intraclass correlation is 0.56, which is in the "weak" range. An examination of the data shows that both panels had very low variance in item ratings (approximately 95% of the items were at DOK level 2). In these cases, Webb recommended using the pairwise agreement to estimate reliability (Webb, 2005). Pairwise agreement for Panel 2, Sub-Study 4 shows very high agreement (0.94).

As may be expected, the interrater agreement is typically higher at the standard (broader) level than at the objective (detail) level.

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 **Mathematics** ***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	4	2	(50%)	2	(50%)	0	-	-	-	**1.50**
1.2	3	1	(33%)	1	(33%)	1	(33%)	-	-	**2.00**
1.3	5	3	(60%)	2	(40%)	0	-	-	-	**1.40**
1.4	2	-	-	2	(100%)	0	-	-	-	**2.00**
1.5	4	1	(25%)	3	(75%)	0	-	-	-	**1.75**
1.6	2	-	-	-	-	2	(100%)	-	-	**3.00**
1 Overall	20	7	(35%)	10	(50%)	3	(15%)	-	-	**1.80**
2.1	6	-	-	6	(100%)	-	-	-	-	**2.00**
2.2	5	1	(20%)	4	(80%)	-	-	-	-	**1.80**
2.3	7	-	-	7	(100%)	-	-	-	-	**2.00**
2 Overall	18	1	(6%)	17	(94%)	-	-	-	-	**1.94**
3.1	4	1	(25%)	3	(75%)	-	-	-	-	**1.75**
3.2	6	1	(17%)	4	(67%)	1	(17%)	-	-	**2.00**
3.3	7	1	(14%)	5	(71%)	1	(14%)	-	-	**2.00**
3.4	8	4	(50%)	4	(50%)	-	-	-	-	**1.50**
3.5	5	-	-	-	-	5	(100%)	-	-	**3.00**
3 Overall	30	7	(23%)	16	(53%)	7	(23%)	-	-	**2.00**
4.1	6	-	-	4	(67%)	2	(33%)	-	-	**2.33**
4.2	7	2	(29%)	4	(57%)	1	(14%)	-	-	**1.86**
4.3	5	-	-	2	(40%)	3	(60%)	-	-	**2.60**
4.4	9	2	(22%)	7	(78%)	-	-	-	-	**1.78**
4.5	5	-	-	2	(40%)	3	(60%)	-	-	**2.60**
4 Overall	32	4	(13%)	19	(59%)	9	(28%)	-	-	**2.16**
5.1	7	-	-	6	(86%)	1	(14%)	-	-	**2.14**
5.2	7	-	-	5	(71%)	2	(29%)	-	-	**2.29**
5.3	7	3	(43%)	4	(57%)	-	-	-	-	**1.57**
5.4	6	3	(50%)	2	(33%)	1	(17%)	-	-	**1.67**
5.5	3	-	-	1	(33%)	2	(67%)	-	-	**2.67**
5 Overall	30	6	(20%)	18	(60%)	6	(20%)	-	-	**2.00**
Total	**130**	**25**	**(19%)**	**80**	**(62%)**	**25**	**(19%)**	**-**	**-**	**2.00**

Table 5 shows that the NAEP objectives associated with Standard 1, "Number Properties and Operations," have an average DOK level of 1.80; the objectives associated with Standard 2, "Measurement," have an average DOK level of 1.94, the objectives associated with Standard 3, "Geometry," have an average DOK level of 2.00; the objectives associated with Standard 4, "Data Analysis, Statistics, and Probability," have an average DOK level of 2.16; and the objectives associated with Standard 5, "Algebra," have an average DOK level of 2.00. The table also shows that the range of DOK levels for the NAEP standards and objectives is 1 to 3.

Table 6 shows the DOK data for the WorkKeys standards.

Table 6: DOK data for the WorkKeys **Applied Mathematics** ***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	4	3 (75%)	1 (25%)	- -	- -	**1.20**
4	7	5 (71%)	2 (29%)	- -	- -	**1.29**
5	7	5 (71%)	2 (29%)	- -	- -	**1.29**
6	9	3 (33%)	5 (56%)	1 (11%)	- -	**1.78**
7	7	- -	6 (86%)	1 (14%)	- -	**2.14**
ALL	**34**	**16 (47%)**	**16 (47%)**	**2 (6%)**	**- -**	**1.58**

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 1.4 to 3 for the NAEP assessment, and from 1.20 to 2.14 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 2.00 and the average WorkKeys standard DOK level being 1.58.

A factor in the difference in the average DOK level of the test standards is suggested by the blueprint analysis. The NAEP assessment is designed to measure twelfth-graders' skill in mathematics content areas and applications that are considered to be grade twelve-appropriate. In contrast, the WorkKeys math assessment is workplace focused and designed to identify individuals' skill level in applying foundational math skills in workplace contexts within a range of mathematical complexity represented in jobs. The focus of the WorkKeys assessment is on problem-solving using foundational math skills, rather than on measuring skill in content areas of high school math.

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 in the appendices show that Panel 1 and Panel 2 members assigned DOK levels to the NAEP items such that the average DOK for all NAEP items considered together was 1.90 for each panel separately and for the two panels combined. The DOK levels assigned to the NAEP items ranged from 1 to 3; no items were labeled DOK Level 4.

For the WorkKeys items, the results for the two panels were also very close, with the vast majority of the items being labeled as DOK Level 2 by most panelists. Just a small number of items received labels of DOK Level 1 or Level 3; no items were labeled DOK Level 4. Both panels assigned DOK levels to the WorkKeys items such that the average DOK for all WorkKeys items considered together was 1.97.

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	1.90	1.97

The average item DOK levels were similar for the two tests, with the WorkKeys average slightly higher than the NAEP average. Table 8 shows the average DOK levels of each item type used on the two tests. For the NAEP items, the two types of constructed-response items (39% of the NAEP item pool) had higher average DOK levels than the NAEP multiple-choice items. However, the majority of all NAEP item types had a DOK level of 2. All WorkKeys items were multiple choice, and all but one had a DOK level of 2.

Table 8: DOK levels by item type

Grade 12 NAEP *Mathematics* 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	17	90	1	0	1.85	108
Short constructed response	5	34	12	0	2.14	51
Extended constructed response	3	14	2	0	1.95	19
WorkKeys *Applied Mathematics* 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	1	57	0	0	1.97	58

** 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, the two tests have opposite results: The average DOK level of the NAEP test standards is higher than the average DOK level of the NAEP test items, while the average DOK level of the WorkKeys test standards is lower than the average DOK level of the WorkKeys test items.

Specifically, for the NAEP assessment, the average DOK level for the standards was 2.00, and the average DOK level for the items was 1.90. The difference between average NAEP standard DOK level and average NAEP item DOK level was 0.10, with the average DOK level of the standards being higher than that of the items.

For the WorkKeys assessment, the average DOK level for the standards was 1.58, and the average DOK level for the items was 1.97. The difference between the average WorkKeys standard DOK level and the average WorkKeys item DOK level was 0.39, with average item DOK level being higher than the average standards DOK level.

Results by Sub-Study

The results of each sub-study are in the next sections (Sub-Study 1, NAEP Grade 12 *Mathematics* items to NAEP Grade 12 *Mathematics* standards; Sub-Study 2, WorkKeys *Applied Mathematics* items to NAEP Grade 12 *Mathematics* items; Sub-Study 3, NAEP Grade 12 *Mathematics* items to WorkKeys *Applied Mathematics* standards; and Sub-Study 4, WorkKeys *Applied Mathematics* items to WorkKeys *Applied Mathematics* 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 *Mathematics* Items to NAEP Grade 12 *Mathematics* 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 *Mathematics* items and the NAEP standards, eight of 178 NAEP test items (4.49%) 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 strongly ("Yes"), weakly ("Weak"), or not at all ("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," and 5 or fewer for "No;" there is no "Weak" value used for this criterion.
- 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 **Mathematics** *items to NAEP Grade 12* **Mathematics** *standards*

NAEP *Mathematics* Standards	Sub-Study 1 — Panels 1 and 2 NAEP Grade 12 *Mathematics* 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) Number properties and operations	25.57	23.62	89%	92%	68	66	0.73	0.74
2) Measurement	19	21	88%	91%	56	63	0.79	0.75
3) Geometry	32.86	33.38	88%	89%	67	71	0.73	0.73
4) Data analysis, statistics, and probability	43	43.25	66%	65%	71	72	0.75	0.76
5) Algebra	58.14	58.75	76%	71%	84	85	0.73	0.73

Table 9 shows 40 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 all 40 points (100%)
- Weak alignment at 0 of 40 points (0%)
- No alignment at 0 of 40 points (0%)

Categorical Concurrence, Depth-of-Knowledge Consistency, Range of Knowledge, and Balance of Representation:

All four alignment criteria of Categorical Concurrence, Depth-of-Knowledge Consistency, Range of Knowledge, and Balance of Representation were met for all five standards. In other words, there were six or more items that targeted each of the five standards, the majority of those items were at or above the DOK levels of the objectives to which they were coded, at least 50% of all objectives within each standard were hit by at least one item, and the number of hits was fairly evenly distributed across the objectives hit. (Note: A "hit" is defined as one reviewer coding an item to an objective.)

The standards for which the Depth-of-Knowledge Consistency percentage was highest were Standard 1, Number properties and operations, and Standard 2, Measurement. The standards that had the highest percentage of objectives hit by items were Standard 4, Data analysis, statistic, and probability, and Standard 5, Algebra.

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 15 reviewers — NAEP Grade 12* Mathematics *items to NAEP Grade 12* Mathematics *standards

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
1	1.1.d	2.57	1%	2.63	1%
	1.1.f	1.29	1%	1.00	1%
	1.1.g	0.29	0%	0.00	0%
	1.1.i	1.14	1%	1.00	1%
	1.2	0.29	0%	0.13	0%
	1.2.b	0.00	0%	0.00	0%
	1.2.c	0.00	0%	0.00	0%
	1.2.d	0.71	0%	0.63	0%
	1.3.a	1.14	1%	1.13	1%
	1.3.b	3.14	2%	2.88	2%
	1.3.c	0.71	0%	0.75	0%
	1.3.d	1.00	1%	0.88	0%
	1.3.f	6.14	3%	5.13	3%
	1.4	0.14	0%	0.00	0%
	1.4.c	1.71	1%	2.50	1%
	1.4.d	0.86	0%	1.00	1%
	1.5.c	2.00	1%	1.63	1%
	1.5.d	1.00	1%	1.00	1%
	1.5.e	0.14	0%	0.25	0%
	1.5.f	0.29	0%	0.00	0%
	1.6.a	0.14	0%	0.13	0%

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
2	2.1.b	0.57	0%	0.88	0%
	2.1.c	0.00	0%	0.00	0%
	2.1.d	0.29	0%	0.25	0%
	2.1.f	3.86	2%	3.75	2%
	2.1.h	2.00	1%	1.88	1%
	2.1.i	1.86	1%	2.25	1%
	2.2.a	0.86	0%	0.88	0%
	2.2.b	1.71	1%	2.25	1%
	2.2.d	0.14	0%	0.13	0%
	2.2.e	0.00	0%	0.63	0%
	2.2.f	2.57	1%	3.25	2%
	2.3.a	0.00	0%	0.38	0%
	2.3.b	0.00	0%	0.13	0%
	2.3.c	2.71	2%	2.00	1%
	2.3.d	0.71	0%	0.50	0%
	2.3.e	0.71	0%	0.75	0%
	2.3.f	1.00	1%	1.00	1%
	2.3.g	0.00	0%	0.13	0%

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
3	3.1.c	0.14	0%	0.00	0%
	3.1.d	0.71	0%	0.88	0%
	3.1.e	1.14	1%	1.13	1%
	3.1.f	0.29	0%	0.50	0%
	3.2.a	0.86	0%	1.00	1%
	3.2.b	0.14	0%	0.13	0%
	3.2.c	1.29	1%	1.25	1%
	3.2.d	1.29	1%	1.00	1%
	3.2.e	0.00	0%	0.00	0%
	3.2.g	1.43	1%	1.75	1%
	3.3.b	6.43	4%	7.00	4%
	3.3.c	0.00	0%	0.00	0%
	3.3.d	1.14	1%	1.00	1%
	3.3.e	0.71	0%	1.25	1%
	3.3.f	0.86	0%	1.00	1%
	3.3.g	2.57	1%	1.75	1%
	3.3.h	1.86	1%	1.88	1%
	3.4	0.14	0%	0.00	0%
	3.4.a	3.00	2%	3.00	2%
	3.4.b	1.29	1%	1.13	1%
	3.4.c	1.14	1%	1.00	1%
	3.4.d	0.00	0%	0.13	0%
	3.4.e	0.86	0%	0.63	0%
	3.4.f	0.71	0%	1.00	1%
	3.4.g	1.00	1%	1.00	1%
	3.4.h	0.71	0%	1.00	1%
	3.5.a	0.29	0%	0.50	0%
	3.5.b	0.00	0%	0.00	0%
	3.5.c	1.00	1%	0.63	0%
	3.5.d	0.86	0%	1.00	1%
	3.5.e	1.00	1%	0.88	0%

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
4	4.1	0.71	0%	1.00	1%
	4.1.a	5.00	3%	5.00	3%
	4.1.b	2.00	1%	2.13	1%
	4.1.c	0.00	0%	0.38	0%
	4.1.d	0.00	0%	0.13	0%
	4.1.e	1.43	1%	1.13	1%
	4.1.f	1.71	1%	1.75	1%
	4.2.a	4.14	2%	3.75	2%
	4.2.b	1.00	1%	0.88	0%
	4.2.c	1.00	1%	0.88	0%
	4.2.d	0.43	0%	0.75	0%
	4.2.e	2.86	2%	2.75	2%
	4.2.f	1.57	1%	1.50	1%
	4.2.g	2.29	1%	2.63	1%
	4.3.a	0.43	0%	0.25	0%
	4.3.b	0.71	0%	1.00	1%
	4.3.c	0.71	0%	0.75	0%
	4.3.d	0.86	0%	0.75	0%
	4.3.e	0.00	0%	0.00	0%
	4.4	0.14	0%	0.00	0%
	4.4.a	1.29	1%	1.00	1%
	4.4.b	1.71	1%	2.00	1%
	4.4.c	3.71	2%	3.88	2%
	4.4.d	0.86	0%	0.75	0%
	4.4.e	1.57	1%	1.38	1%
	4.4.h	2.29	1%	2.13	1%
	4.4.i	0.14	0%	0.50	0%
	4.4.j	1.43	1%	1.38	1%
	4.4.k	0.00	0%	0.00	0%
	4.5.a	1.00	1%	1.00	1%
	4.5.b	1.00	1%	0.75	0%
	4.5.c	0.00	0%	0.00	0%
	4.5.d	0.00	0%	0.00	0%
	4.5.e	1.00	1%	1.13	1%
	4.1	0.71	0%	1.00	1%

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
5	5.1.a	1.86	1%	1.75	1%
	5.1.b	2.00	1%	2.25	1%
	5.1.e	5.29	3%	4.50	3%
	5.1.g	1.00	1%	0.88	0%
	5.1.h	0.14	0%	0.13	0%
	5.1.i	1.00	1%	1.00	1%
	5.1.j	1.14	1%	1.00	1%
	5.2.a	1.71	1%	1.88	1%
	5.2.b	1.14	1%	1.38	1%
	5.2.d	0.86	0%	0.75	0%
	5.2.e	0.00	0%	0.00	0%
	5.2.f	2.00	1%	2.13	1%
	5.2.g	0.86	0%	1.00	1%
	5.2.h	1.00	1%	1.00	1%
	5.3.b	3.71	2%	3.75	2%
	5.3.c	4.43	2%	3.25	2%
	5.3.d	2.14	1%	3.38	2%
	5.3.e	3.14	2%	3.25	2%
	5.3.f	4.71	3%	4.50	3%
	5.3.g	1.29	1%	1.00	1%
	5.3.h	0.00	0%	0.25	0%
	5.4.a	3.14	2%	3.38	2%
	5.4.c	5.43	3%	5.50	3%
	5.4.d	1.00	1%	1.25	1%
	5.4.e	2.00	1%	2.25	1%
	5.4.f	1.71	1%	2.00	1%
	5.4.g	0.57	0%	0.63	0%
	5.5	0.00	0%	0.13	0%
	5.5.a	0.29	0%	0.00	0%
	5.5.b	3.71	2%	3.63	2%
	5.5.c	0.86	0%	1.00	1%

There are nearly as many objectives as there are items in the pool used for this study. In addition, the panelists did not judge any items to be uncodable. Therefore, if all objectives are to be covered by the test items, then no objective could have more than two items aligned to it.

Both panels coded at least one NAEP item to 82% of the NAEP objectives (not including "generic" objectives). One or both panels coded no items to 18% of the objectives. The objectives to which one or both panels coded no items are shown in the following list. According to the NAEP mathematics framework, objectives marked with an asterisk represent content that is beyond the typical three-year mathematics course of study in high school and is, therefore, selected for inclusion less often on the assessment. Four of the 23 objectives in the list are marked with an asterisk.

- 1.1.g — "Represent, interpret, or compare expressions or problem situations involving absolute values."
- 1.2.b — "Identify situations where estimation is appropriate, determine the needed degree of accuracy, and analyze *the effect of the estimation method on the accuracy of results."
- 1.2.c — "Verify solutions or determine the reasonableness of results in a variety of situations."
- 1.5.f — "Recognize properties of the number system (whole numbers, integers, rational numbers, real numbers, and complex numbers) and how they are related to each other, and identify examples of each type of number."
- 2.1.c — "Estimate or compare perimeters of areas of two-dimensional geometric figures."
- 2.2.e — "Determine appropriate accuracy of measurement in problem situations (e.g., the accuracy of measurement of the dimensions to obtain a specified accuracy of area) and find the measure to that degree of accuracy."
- 2.3.a — "Solve problems involving indirect measurement."
- 2.3.b — "Solve problems using the fact that trigonometric rations (sine, cosine, and tangent) stay constant in similar triangles."
- 2.3.g — "*Use the law of cosines and the law of sines to find unknown sides and angles of a triangle."
- 3.1.c — "Give precise mathematical descriptions or definitions of geometric shapes in the plane and in three-dimensional space."
- 3.2.e — "Justify relationships of congruence and similarity and apply these relationships using scaling and proportional reasoning."
- 3.3.c — "Represent problem situations with geometric models to solve mathematical or real-world problems."
- 3.4.d — "Represent two-dimensional figures algebraically using coordinates and/or equations."
- 3.5.b — "Determine the role of hypotheses, logical implications, and conclusion in proofs of geometric theorems."
- 4.1.c — "Solve problems involving univariate or bivariate data."
- 4.1.d — "Given a graphical or tabular representation of a set of data, determine whether information is represented effectively and appropriately."
- 4.3.e —"* Recognize the differences in design and in conclusions between randomized experiments and observational studies."
- 4.4.k — "*Use the binomial theorem to solve problems."
- 4.5.c — "*Recognize, use, and distinguish between the processes of mathematical (deterministic) and statistical modeling."
- 4.5.d — "Recognize when arguments based on data confuse correlation with causation."
- 5.2.e — "Make inferences or predictions using an algebraic model of a situation."
- 5.3.h — "Use basic properties of exponents and *logarithms to solve problems."
- 5.5.a — "Use algebraic properties to develop a valid mathematical argument."

Sub-Study 2: WorkKeys *Applied Mathematics* Items to NAEP Grade 12 *Mathematics* Standards

In Sub-Study 2, the alignment between the WorkKeys *Applied Mathematics* items and the NAEP Grade 12 *Mathematics* standards, two items were coded to a generic standards by one panelist; however 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 strongly ("Yes"), weakly ("Weak"), or not at all ("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," and 5 or fewer for "No;" there is no "Weak" value used for this criterion.
- 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* Applied Mathematics *items to NAEP Grade 12* Mathematics *standards

NAEP *Mathematics* Standards	Sub-Study 2 — Panels 1 and 2 WorkKeys *Applied Mathematics* 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) Number properties and operations	41	39.5	99%	99%	17**	17**	0.55**	0.62*
2) Measurement	17.57	18	100%	100%	28**	29**	0.61*	0.60*
3) Geometry	0.14**	0**	100%	0%**	0**	0**	0.14**	0**
4) Data analysis, statistics, and probability	3.14**	4.12**	67%	73%	4**	5**	0.92	0.89
5) Algebra	0**	0**	0%**	0%**	0**	0**	0**	0**

Table 11 shows 40 points for which the degree of alignment between the WorkKeys *Applied Mathematics* items and the NAEP Grade 12 *Mathematics* standards is calculated. The table shows that the panels' judgment resulted in the following:

- Alignment criteria met at 13 of 40 points (32.5%)
- Weak alignment at 3 of 40 points (7.5%)
- No alignment at 24 of 40 points (60%)

The data for Standard 1, "Number properties and operations," and Standard 2, "Measurement," show alignment using the Categorical Concurrence and Depth-of-Knowledge Consistency criteria, weak alignment using the Balance of Representation criterion, and no alignment using the Range of Knowledge Criterion. The data indicate alignment for Standard 4, "Data analysis, statistics, and probability," only using the DOK Consistency and Balance of Representation criteria. No items were coded to Standard 5, "Algebra," and only one panelist coded one item to an objective within Standard 3, "Geometry," so there is no alignment to these standards.

Categorical Concurrence:

The Categorical Concurrence criterion was met (more than six items coded to objectives within the standard) for Standards 1, "Number Properties and Operations," and 2, "Measurement." The WorkKeys items coded to these standards covered some objectives within the standards but not all.

Just a few items were coded to Standard 4, "Data Analysis, Statistics, and Probability," which expects students to use a wide variety of statistical techniques for all phases of the data analysis process. These items were coded to Objective 4.1.a, "Read or interpret graphical or tabular representations of data," or to Objective 4.2.a, "Calculate, interpret or use summary statistics for distributions of data including measures of typical value (mean, median), position (quartiles, percentiles), and spread (range, interquartile range, variance, and standard deviation)." Again, the WorkKeys items coded to these objectives covered some but not all of the content represented by the objectives.

Just one panelist coded one item to a standard within Standard 3, "Geometry," which expects students to be able to represent geometric transformations algebraically. Additionally, no items were coded to Standard 5, "Algebra,"

Depth-of-Knowledge Consistency:

For the NAEP standards to which the panels coded WorkKeys items, there was strong DOK consistency. The exception is for Standard 3, "Geometry." Here, the DOK Consistency criterion is met for Panel 1, but is not met for Panel 2. As noted above, the reason for this is that just one member of Panel 1 coded one item to one objective in Standard 3; no other panelists coded any items to objectives within Standard 3. The panelist coded this one item to the same DOK level as the objective (Level 2); therefore the DOK Consistency is 100%. Because Panel 2 did not code any items to this standard, the DOK Consistency for that panel is 0%. The few items that were coded to objectives within Standard 4 were, in general, at the same DOK level as the objective.

Looking more closely at how the NAEP 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 15 reviewers — WorkKeys **Applied Mathematics** ***Items to NAEP Grade 12*** **Mathematics** ***standards***

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
1	1.1.d	0.00	0%	0.00	0%
	1.1.f	0.00	0%	0.00	0%
	1.1.g	0.00	0%	0.00	0%
	1.1.i	0.00	0%	0.00	0%
	1.2.b	0.00	0%	0.00	0%
	1.2.c	0.00	0%	0.00	0%
	1.2.d	0.00	0%	0.00	0%
	1.3.a	0.00	0%	0.00	0%
	1.3.b	0.29	0%	0.75	1%
	1.3.c	0.00	0%	0.00	0%
	1.3.d	0.00	0%	0.00	0%
	1.3.f	31.14	50%	27.00	44%
	1.4	0.29	0%	0.00	0%
	1.4.c	5.57	9%	6.63	11%
	1.4.d	3.57	6%	5.13	8%
	1.5.c	0.00	0%	0.00	0%
	1.5.d	0.14	0%	0.00	0%
	1.5.e	0.00	0%	0.00	0%
	1.5.f	0.00	0%	0.00	0%
	1.6.a	0.00	0%	0.00	0%
	1.6.b	0.00	0%	0.00	0%
2	2.1.b	0.00	0%	0.00	0%
	2.1.c	0.00	0%	0.13	0%
	2.1.d	0.00	0%	0.00	0%
	2.1.f	6.71	11%	7.13	12%
	2.1.h	1.00	2%	1.00	2%
	2.1.i	2.29	4%	1.38	2%
	2.2.a	0.00	0%	0.00	0%
	2.2.b	6.57	11%	7.13	12%
	2.2.d	0.00	0%	0.00	0%
	2.2.e	0.00	0%	0.13	0%
	2.2.f	1.00	2%	1.00	2%
	2.3.a	0.00	0%	0.13	0%
	2.3.b	0.00	0%	0.00	0%
	2.3.c	0.00	0%	0.00	0%
	2.3.d	0.00	0%	0.00	0%
	2.3.e	0.00	0%	0.00	0%
	2.3.f	0.00	0%	0.00	0%
	2.3.g	0.00	0%	0.00	0%

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
3	3.1.c	0.00	0%	0.00	0%
	3.1.d	0.00	0%	0.00	0%
	3.1.e	0.00	0%	0.00	0%
	3.1.f	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%
	3.2.d	0.00	0%	0.00	0%
	3.2.e	0.00	0%	0.00	0%
	3.2.g	0.00	0%	0.00	0%
	3.3.b	0.14	0%	0.00	0%
	3.3.c	0.00	0%	0.00	0%
	3.3.d	0.00	0%	0.00	0%
	3.3.e	0.00	0%	0.00	0%
	3.3.f	0.00	0%	0.00	0%
	3.3.g	0.00	0%	0.00	0%
	3.3.h	0.00	0%	0.00	0%
	3.4.a	0.00	0%	0.00	0%
	3.4.b	0.00	0%	0.00	0%
	3.4.c	0.00	0%	0.00	0%
	3.4.d	0.00	0%	0.00	0%
	3.4.e	0.00	0%	0.00	0%
	3.4.f	0.00	0%	0.00	0%
	3.4.g	0.00	0%	0.00	0%
	3.4.h	0.00	0%	0.00	0%
	3.5.a	0.00	0%	0.00	0%
	3.5.b	0.00	0%	0.00	0%
	3.5.c	0.00	0%	0.00	0%
	3.5.d	0.00	0%	0.00	0%
	3.5.e	0.00	0%	0.00	0%

NAEP Standards	**Objectives**	**Panel 1**		**Panel 2**	
		Mean Hits	**% of Total Hits**	**Mean Hits**	**% of Total Hits**
4	4.1.a	0.43	1%	1.00	2%
	4.1.b	0.00	0%	0.00	0%
	4.1.c	0.00	0%	0.00	0%
	4.1.d	0.00	0%	0.00	0%
	4.1.e	0.00	0%	0.00	0%
	4.1.f	0.00	0%	0.00	0%
	4.2.a	2.71	4%	3.00	5%
	4.2.b	0.00	0%	0.00	0%
	4.2.c	0.00	0%	0.00	0%
	4.2.d	0.00	0%	0.00	0%
	4.2.e	0.00	0%	0.00	0%
	4.2.f	0.00	0%	0.00	0%
	4.2.g	0.00	0%	0.00	0%
	4.3.a	0.00	0%	0.00	0%
	4.3.b	0.00	0%	0.00	0%
	4.3.c	0.00	0%	0.00	0%
	4.3.d	0.00	0%	0.00	0%
	4.3.e	0.00	0%	0.00	0%
	4.4.a	0.00	0%	0.00	0%
	4.4.b	0.00	0%	0.00	0%
	4.4.c	0.00	0%	0.00	0%
	4.4.d	0.00	0%	0.00	0%
	4.4.e	0.00	0%	0.00	0%
	4.4.h	0.00	0%	0.00	0%
	4.4.i	0.00	0%	0.00	0%
	4.4.j	0.00	0%	0.13	0%
	4.4.k	0.00	0%	0.00	0%
	4.5.a	0.00	0%	0.00	0%
	4.5.b	0.00	0%	0.00	0%
	4.5.c	0.00	0%	0.00	0%
	4.5.d	0.00	0%	0.00	0%
	4.5.e	0.00	0%	0.00	0%

NAEP Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
5	5.1.a	0.00	0%	0.00	0%
	5.1.b	0.00	0%	0.00	0%
	5.1.e	0.00	0%	0.00	0%
	5.1.g	0.00	0%	0.00	0%
	5.1.h	0.00	0%	0.00	0%
	5.1.i	0.00	0%	0.00	0%
	5.1.j	0.00	0%	0.00	0%
	5.2.a	0.00	0%	0.00	0%
	5.2.b	0.00	0%	0.00	0%
	5.2.d	0.00	0%	0.00	0%
	5.2.e	0.00	0%	0.00	0%
	5.2.f	0.00	0%	0.00	0%
	5.2.g	0.00	0%	0.00	0%
	5.2.h	0.00	0%	0.00	0%
	5.3.b	0.00	0%	0.00	0%
	5.3.c	0.00	0%	0.00	0%
	5.3.d	0.00	0%	0.00	0%
	5.3.e	0.00	0%	0.00	0%
	5.3.f	0.00	0%	0.00	0%
	5.3.g	0.00	0%	0.00	0%
	5.3.h	0.00	0%	0.00	0%
	5.4.a	0.00	0%	0.00	0%
	5.4.c	0.00	0%	0.00	0%
	5.4.d	0.00	0%	0.00	0%
	5.1.a	0.00	0%	0.00	0%
	5.1.b	0.00	0%	0.00	0%
	5.1.e	0.00	0%	0.00	0%
	5.1.g	0.00	0%	0.00	0%
	5.1.h	0.00	0%	0.00	0%
	5.1.i	0.00	0%	0.00	0%

Range of Knowledge:

Table 12 illustrates why the Range of Knowledge and Balance of Representation alignment criteria were largely unmet for all five NAEP standards. In short, there was a significant number of objectives to which no WorkKeys items were coded, and Standards 3 and 5 ("Geometry" and "Algebra"), respectively, had only one or no items coded to them. Thus, all WorkKeys items were clustered around a somewhat limited number of objectives, resulting in a limited range.

Panel 1 coded the WorkKeys items such that the Balance of Representation criterion was not met for Standard 1 ("Number Properties and Operations"), whereas Panel 2 coded the items such that the criterion was weakly met. In both panels, most of the items coded to Standard 1 targeted 1.3.f, 1.4.c, and 1.4.d. In Panel 2, there were several more instances of one panelist coding an item to

1.3.b than there were in Panel 1, and this was enough to move the Balance of Representation for Panel 2 to “weak.”

Another significant factor contributing to why the WorkKeys items did not meet these two criteria, as well as Categorical Concurrence, is that there were only 58 unique WorkKeys items considered for the study, and there were 154 NAEP objectives, including “generics,” to which the items were to be coded. Thus, it was impossible for the WorkKeys items to cover all of the NAEP objectives without coding WorkKeys items to multiple NAEP objectives. If a larger number of items from the WorkKeys item pool of hundreds of math items had been included in the study, some additional objectives may have been covered. However, as the blueprint analysis indicates, there are also significant portions of the math content areas included on the NAEP assessment that are not part of the WorkKeys blueprint. Therefore, the fact that there were only slightly more than a third as many WorkKeys items as NAEP objectives is not the only reason the NAEP objectives were not fully covered by the WorkKeys items.

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 alignment is strong 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 2. The WAT calculations indicate strong Balance of Representation alignment index values — values of 0.92 and 0.89 for Standard 4, “Data Analysis, Statistics, and Probability.” It might be inferred from these data that the WorkKeys items are, for the most part, coded to the NAEP objectives within Standard 4 fairly evenly. However, as discussed earlier in this section, there were not many items coded to this standard, and, furthermore, only two objectives within the standard were targeted. Thus, the WorkKeys items are not coded evenly to the NAEP objectives within Standard 4, and Table 12 helps to illustrate this.

In summary, Table 12 shows that both panels coded at least one WorkKeys item to 8.5% of the NAEP objectives (not including “generic” objectives). One or both panels coded no items to 91.5% of the objectives, which are not listed here due to their high number (refer to Appendix E for the list of objectives). The NAEP objectives targeted by the most WorkKeys items include problem-solving applications of number operations and measurement. The NAEP objectives to which no WorkKeys items aligned are most commonly found in Standards 3 (“Geometry”) and 5 (“Algebra”).

Sub-Study 3: NAEP Grade 12 *Mathematics* Items to WorkKeys *Applied Mathematics* Standards

In Sub-Study 3, the alignment between the NAEP Grade 12 *Mathematics* items and the WorkKeys *Applied Mathematics* standards, 148 of 178 NAEP items (83.15%) were rated as uncodable to WorkKeys standards by at least one panelist. Of these, 89 items (50%) were deemed uncodable 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 strongly ("Yes"), weakly ("Weak"), or not at all ("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," and 5 or fewer for "No;" there is no "Weak" value used for this criterion.
- 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* Mathematics *items to WorkKeys* Applied Mathematics *standards

WorkKeys *Applied Mathematics* Standards	**Sub-Study 3 — Panels 1 and 2 NAEP Grade 12 *Mathematics* 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) A single type of basic mathematics operation; no reordering or extraneous information	0.29**	0.25**	100%	100%	7**	6**	0.29**	0.25**
4) Multiple types of mathematical operations; reordering, extraneous information	12.71	14	97%	88%	29**	34**	0.78	0.70
5) Application of logic and calculation; conversions	11	10.62	91%	96%	43*	46*	0.77	0.75
6) Complex and multiple-step calculations; manipulating formulas	10	10.5	100%	81%	62	64	0.74	0.74
7) Nonlinear functions, complex calculations and conversions	22.57	24.75	94%	92%	47*	46*	0.43**	0.48**

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 26 of 40 points (65%)
- Weak alignment at 4 of 40 points (10%)
- No alignment at 10 of 40 points (25%)

For the alignment criteria of Categorical Concurrence and Depth-of-Knowledge Consistency, the data show that both panels judged the alignment criteria to be met, with one exception.

Categorical Concurrence:
The data show strong alignment for Categorical Concurrence for four of the five WorkKeys standards. There is not alignment at Standard 3, "A single type of basic mathematics operation; no reordering or extraneous information."

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. Eighty-nine 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* Mathematics *items to WorkKeys* Applied Mathematics *standards

	Panel 1	Panel 2
Codable items	60	89
Uncodable items	118	89
Total assessment items	178	178

All of the 89 items unanimously deemed uncodable by Panel 2 are included in the 118 unanimously deemed uncodable by Panel 1. The remaining 29 items unanimously deemed uncodable by Panel 1 were also deemed uncodable by the majority of panelists in Panel 2; however, the judgment of Panel 2 members was not unanimous for these 29 items, and in many cases there was just one panelist who felt a given item was codable while all the rest of the panelists felt it was uncodable. Thus, the judgments of the two panels are closer than these two numbers may seem to indicate.

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 15 reviewers — NAEP Grade 12* Mathematics *items to WorkKeys* Applied Mathematics *standards

	Panel 1		Panel 2	
	Mean Hits	Percentage	Mean Hits	Percentage
Codable	56.57	31.78%	60.12	33.78%
Uncodable	121.43	68.22%	117.88	66.22%
Total	178		178	

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 15 reviewers — NAEP Grade 12 **Mathematics** ***items to WorkKeys*** **Applied Mathematics** ***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	0.29	0.51%	0.16%	0.25	0.42%	0.14%
4	12.71	22.47%	7.14%	14	23.29%	7.87%
5	11	19.44%	6.18%	10.62	17.66%	5.97%
6	10	17.68%	5.62%	10.50	17.47%	5.90%
7	22.57	39.90%	12.68%	24.75	41.17%	13.90%
Total	56.57	100.00%	31.78%	60.12	100.00%	33.78%

Thus, it can be seen that roughly a third to one-half of the NAEP items to be codable to the WorkKeys standards. Furthermore, Standard 7 received the greatest percentage of hits, followed by Standards 4, then 5, then 6, and then 3.

Depth-of-Knowledge Consistency:
The data show that the alignment criterion of Depth-of-Knowledge Consistency was met at all standards in the judgment of both panels. This is to be expected, given the average DOK level of the WorkKeys standards is 1.58 and the average DOK level of the NAEP items is 1.90.

In contrast to the alignment criteria of 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 criteria, which are 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 met the threshold values for the Range of Knowledge alignment criterion at Standard 6 only. The criterion was weakly met for Standards 5 and 7. It was not met for Standards 3 or 4.

Standard 6: This standard reads, “Complex and multiple-step calculations; manipulating formulas.” The panelists aligned NAEP items to seven of the nine objectives within this standard.

Standards 5 and 7: Standard 5 reads, “Application of logic and calculation; conversions,” and Standard 7 reads, “Nonlinear functions, complex calculations and conversions.” 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. For Standard 5, Panel 1 aligned no items to four of the seven objectives, while Panel 2 aligned no items to two of the seven objectives and a few individuals aligned an item with two other objectives. For Standard 7, two of the seven objectives were not hit by either panel, and two other objectives received a minimal number of hits.

Standards 3 and 4: Standard 3 reads, “A single type of basic mathematics operation; no reordering or extraneous information,” and Standard 4 reads, “Multiple types of mathematical operations; reordering, extraneous information.” For Standard 3, four panelists coded three items to two different objectives; no other items were coded to objectives within this standard. For Standard 4, five of seven objectives received no or virtually no hits by any NAEP items.

Thus, while the blueprint analysis showed that the content included in the NAEP framework is broader than that included in the WorkKeys framework, there are still numerous areas of the WorkKeys framework that are not covered by the NAEP items. A list of the areas not covered is included later in this section; they are primarily workplace applications of math concepts.

Balance of Representation:
As discussed in the earlier section on Sub-Study 2, Balance of Representation is defined this way: “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.” However, as discussed with respect to Sub-Study 2, 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. Thus, the balance index indicates how the items coded to a standard are distributed among the objectives to which they were coded, not how they are distributed among all objectives within the standard.

For Sub-Study 3, the WAT calculations indicate strong Balance of Representation alignment index values — values ranging from 0.70 to 0.78 — for six of the ten points in the sub-study. The other four points show Balance of Representation alignment index values of 0.25 to 0.48, which put them in the “no alignment” range. It might be inferred from these data that the NAEP items are, for the most part, 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 15 reviewers — NAEP Grade 12 **Mathematics** ***items to WorkKeys*** **Applied Mathematics** ***standards***

WorkKeys Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
3	3.1	0.29	1%	0.13	0%
	3.2	0.00	0%	0.13	0%
	3.3	0.00	0%	0.00	0%
	3.4	0.00	0%	0.00	0%

WorkKeys Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
4	4.1	9.14	16%	10.13	17%
	4.2	0.00	0%	0.00	0%
	4.3	3.57	6%	3.50	6%
	4.4	0.00	0%	0.00	0%
	4.5	0.00	0%	0.00	0%
	4.6	0.00	0%	0.13	0%
	4.7	0.00	0%	0.25	0%
5	5.1	6.00	11%	5.13	9%
	5.2	1.29	2%	1.13	2%
	5.3	0.00	0%	0.00	0%
	5.4	0.00	0%	0.00	0%
	5.5	0.00	0%	0.13	0%
	5.6	3.71	7%	4.13	7%
	5.7	0.00	0%	0.13	0%
6	6.1	3.14	6%	3.00	5%
	6.2	0.86	2%	0.88	1%
	6.3	0.71	1%	0.88	1%
	6.4	1.14	2%	1.00	2%
	6.5	0.00	0%	0.00	0%
	6.6	0.00	0%	0.00	0%
	6.7	3.14	6%	3.38	6%
	6.8	1.00	2%	1.13	2%
	6.9	0.00	0%	0.25	0%
7	7.1	1.43	3%	2.63	4%
	7.2	0.00	0%	0.00	0%
	7.3	0.00	0%	0.00	0%
	7.4	0.29	1%	0.13	0%
	7.5	0.00	0%	0.50	1%
	7.6	1.00	2%	0.88	1%
	7.7	19.86	35%	20.63	34%

Thus, this table illustrates that, when calculating Balance of Representation, the WAT calculations consider only the objectives to which at least one item has been coded. For instance, at Standard 4, a significant number of NAEP items were coded to each of Objectives 4.1 and 4.3. The other five objectives within Standard 4 received virtually no hits. In Panel 1, Objective 4.1 received about 5% of the coded items and Objective 4.3 received about 2%; in Panel 2, Objective 4.1 received about 7% of the coded items and Objective 4.3 received about 2%. The remaining objectives, using rounded values, received 0% of the coded items. Still, the WAT shows that the Balance of Representation index value is in the “Yes” range, 0.78 for Panel 1 and 0.70 for Panel 2.

To summarize Table 17, both panels coded at least one NAEP item to 47% of the WorkKeys objectives. One or both panels coded no items to 53% of the objectives. The objectives to which one or both panels coded no items are shown in the following list:

- 3.2 — “Add or subtract negative numbers”

- 3.3 — “Change numbers from one form to another using whole numbers, fractions, decimals, or percentages”
- 3.4 — “Convert simple money and time units (e.g., hours to minutes)”
- 4.2 — “Multiply negative numbers”
- 4.4 — “Add commonly known fractions, decimals, or percentages (e.g., 1/2, .75, 25%)”
- 4.5 — “Add up to three fractions that share a common denominator”
- 4.6 — “Multiply a mixed number by a whole number or decimal”
- 4.7 — “Put the information in the right order before performing calculations”
- 5.3 — “Calculate using mixed units (e.g., 3.5 hours and 4 hours 30 minutes)”
- 5.4 — “Divide negative numbers”
- 5.5 — “Find the best deal using one- and two-step calculations and then comparing results”
- 5.7 — “Calculate percent discounts or markups”
- 6.5 — “Find mistakes in questions that belong at Levels 3, 4, and 5”
- 6.6 — “Find the best deal and use the result for another calculation”
- 6.9 — “Calculate multiple rates”
- 7.2 — “Find mistakes in Level 6 questions”
- 7.3 — “Convert between systems of measurement that involve fractions, mixed numbers, decimals, and/or percentages”
- 7.5 — “Set up and manipulate complex ratios or proportions”

The WorkKeys objectives most frequently covered by NAEP items include geometry content; fractions, ratios, percentages, or mixed numbers; and basic statistical concepts. The WorkKeys objectives that were targeted least often by the NAEP items range across a variety of workplace-oriented math skills: conversions, determining the best deal, finding errors, and calculating discounts or markups.

Sub-Study 4: WorkKeys *Applied Mathematics* Items to WorkKeys *Applied Mathematics* Standards

In Sub-Study 4, the alignment between the WorkKeys *Applied Mathematics* items and the WorkKeys *Applied Mathematics* standards, no items were coded to a generic standard or were judged as uncodable by any panelists. This study was conducted remotely in February 2010, with four of the original members from each panel and the original facilitators.

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 strongly (“Yes”), weakly (“Weak”), or not at all (“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,” and 5 or fewer for “No;” there is no “Weak” value used for this criterion.

- 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* Applied Mathematics *items to WorkKeys* Applied Mathematics *standards

WorkKeys *Applied Mathematics* Standards	**Sub-Study 4 — Panels 1 and 2 WorkKeys *Applied Mathematics* 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) A single type of basic mathematics operation; no reordering or extraneous information	10	10.25	95%	98%	25**	25**	1	1
4) Multiple types of mathematical operations; reordering, extraneous information	19.25	19	99%	100%	43*	43*	0.70	0.71
5) Application of logic and calculation; conversions	11.75	11.5	100%	100%	71	71	0.84	0.90
6) Complex and multiple-step calculations; manipulating formulas	16	16.5	100%	98%	67	67	0.76	0.74
7) Nonlinear functions, complex calculations and conversions	2**	1.75**	100%	100%	14**	14**	1	1

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 32 of 40 points (80%)

- Weak alignment at 2 of 40 points (5%)
- No alignment at 6 of 40 points (15%)

Of the four alignment criteria, Range of Knowledge is the criterion with the fewest points of strong alignment.

Categorical Concurrence:
This criterion was met strongly for Standards 3, 4, 5, and 6, but not for Standard 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 7 reads, "Nonlinear functions, complex calculations and conversions." Panelists coded two items to Objective 7.5 only ("Set up and manipulate complex ratios or proportions").

Depth-of-Knowledge Consistency:
Alignment to the Depth-of-Knowledge Consistency criterion was strong ("Yes") for all standards.

Range of Knowledge:
The Range of Knowledge criterion was met for Standards 5 ("Application of logic and calculation; conversions") and 6 ("Complex and multiple-step calculations; manipulating formulas"), and it was weakly met for Standard 4 ("Multiple types of mathematical operations; reordering, extraneous information"). The criterion was not met for Standards 3 ("A single type of basic mathematics operation; no reordering or extraneous information") and 7 ("Nonlinear functions, complex calculations and conversions"). The WAT calculations require that 50% or more of the objectives have items coded to them in order for the criterion to be considered met. In Standard 4, three of the seven objectives had items coded to them, putting the standard just below the threshold for meeting the criterion. Within each of Standards 3 and 7, just one objective was hit. For instance, in Standard 3, Objective 3.1 received about 10% of the total hits, but because the other three objectives within the standard did not receive any hits, the Range of Knowledge criterion was not met for Standard 3.

Balance of Representation:
All five standards met the criterion for alignment in the Balance of Representation category.

However, as discussed previously in this report, the statistics associated with this criterion may be misleading on the surface. For instance, as just described, Standard 3 received about 10% of the total hits for this sub-study, but all hits were targeted at Objective 3.1. Because none of the other three objectives within this standard received any hits, the WAT software shows that the balance index is a perfect 1.

Therefore, it may be more informative to look specifically 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 8 reviewers — WorkKeys* Applied Mathematics *items to WorkKeys* Applied Mathematics *standards

WorkKeys Standards	Objectives	Panel 1		Panel 2	
		Mean Hits	% of Total Hits	Mean Hits	% of Total Hits
3	3.1	10	17%	10.25	18%
	3.2	0	0%	0	0%
	3.3	0	0%	0	0%
	3.4	0	0%	0	0%
4	4.1	12.25	21%	11.75	20%
	4.2	0	0%	0	0%
	4.3	6	10%	6.25	11%
	4.4	0	0%	0	0%
	4.5	1	2%	1	2%
	4.6	0	0%	0	0%
	4.7	0	0%	0	0%
5	5.1	0	0%	0	0%
	5.2	2	3%	1.75	3%
	5.3	3.25	6%	3	5%
	5.4	0	0%	0	0%
	5.5	1.75	3%	2	3%
	5.6	2	3%	2	3%
	5.7	2.75	5%	2.75	5%
6	6.1	3.75	6%	4.5	8%
	6.2	1	2%	1	2%
	6.3	1	2%	1	2%
	6.4	2.25	4%	2	3%
	6.5	0	0%	0	0%
	6.6	3	5%	3	5%
	6.7	5	9%	5	9%
	6.8	0	0%	0	0%
	6.9	0	0%	0	0%
7	7.1	0	0%	0	0%
	7.2	0	0%	0	0%
	7.3	0	0%	0	0%
	7.4	0	0%	0	0%
	7.5	2	3%	1.75	3%
	7.6	0	0%	0	0%
	7.7	0	0%	0	0%

To summarize Table XX, both panels coded at least one WorkKeys item to 47% of the WorkKeys objectives. One or both panels coded no items to 53% of the objectives. The objectives to which one or both panels coded no items are shown in the following list:

- 3.2 — “Add or subtract negative numbers”
- 3.3 — “Change numbers from one form to another using whole numbers, fractions, decimals, or percentages”
- 3.4 — “Convert simple money and time units (e.g., hours to minutes)”

- 4.2 — "Multiply negative numbers"
- 4.4 — "Add commonly known fractions, decimals, or percentages (e.g., 1/2, .75, 25%)"
- 4.6 — "Multiply a mixed number by a whole number or decimal"
- 4.7 — "Put the information in the right order before performing calculations"
- 5.1 — "Decide what information, calculations, or unit conversions to use to solve the problem"
- 5.4 — "Divide negative numbers"
- 6.5 — "Find mistakes in questions that belong at Levels 3, 4, and 5"
- 6.8 — "Find the volume of rectangular solids"
- 6.9 — "Calculate multiple rates"
- 7.1 — "Solve problems that include nonlinear functions and/or that involve more than one unknown"
- 7.2 — "Find mistakes in Level 6 questions"
- 7.3 — "Convert between systems of measurement that involve fractions, mixed numbers, decimals, and/or percentages"
- 7.4 — "Calculate multiple areas and volumes of spheres, cylinders, or cones"
- 7.6 — "Find the best deal when there are several choices"
- 7.7 — "Apply basic statistical concepts"

The WorkKeys objectives targeted most frequently by the items included in the pool for this study include those involving one or two operations; ratios, percentages, or mixed numbers; calculating areas along with unit conversions; and determining the best deal. The objectives targeted least often by the WorkKeys items are those involving negative numbers, finding mistakes, the volume of objects, nonlinear functions, and basic statistical concepts.

It is important to note that the reason some of the objectives were not targeted by items is likely due to item sampling. This study 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 178 items available on the 2009 NAEP assessment were studied, only 58 unique items from the complete pool of hundreds of operational WorkKeys items were used in these studies. 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 alignment according to 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.

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:

- Two of 16 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 3.14 and 3.69 (out of 4).
- Participants wanted to know more about how the NAEP standards might be affected by these studies. Some were concerned about the DOK criterion, specifically whether it was a useful criterion and whether they were applying it correctly.

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, 1/11/10	Tuesday, 1/12/10	Wednesday, 1/13/10	Thursday, 1/14/10
1. How comfortable do you feel with the process of assigning DOK levels? **(Scale = 1 – 4)**	3.36	3..36	3.46	3.45
2. How well did your group facilitator facilitate the consensus process? **(Scale = 1 – 3)**	3.00	3.00	3.00	3.00

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 responses about the consensus process were "Very Well," 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.

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 areas of the standards they felt were not covered well. Some mentioned topics that were assessed at an elementary level and other topics that were overemphasized.

DOK levels: Panelists felt that the majority of items were at DOK Level 2. Some noted the absence of DOK Level 4 items, but concluded it would difficult to assess Level 4 in this format.

Standards: There was a wide range of opinion expressed about whether the standards were written at the appropriate grade level and at an appropriate level of specificity. Panelists' opinions on this issue ranged from "No" to "Yes." Some noted that many standards included two or more statements of skills or knowledge joined by "and" and were therefore broader due to their inclusion of more than one main part.

Alignment: Panel 1 results indicate that 86% felt there was acceptable alignment and 14% felt there needs to be slight improvement to the alignment. Panel 2 results indicate that 75% felt there was acceptable alignment while 25% felt there needs to be slight improvement.

Evaluation of Sub-Study 2 —WorkKeys to NAEP

Coverage of the standards by the items: In general, the WorkKeys items covered Standard 1 ("Number Properties and Operations") and Standard 2 ("Measurement"). Standard 4 ("Data Analysis, Statistics, and Probability") was minimally targeted, while Standard 3 ("Geometry") and Standard 5 ("Algebra") were not covered beyond basic perimeters, areas, and volumes. The emphasis was on computation and less so on mathematical reasoning. Panelists noted specifically a number of areas that were not covered by WorkKeys items. One panelist commented that the general nature of some objectives, such as Objective 1.3.f, "Solve application problems involving numbers, including rational and common irrationals," allowed the coding of many items that required computation in a real-world context.

DOK levels: Panelists pointed out that many WorkKeys items had a DOK level of 2. Some commented that few items were rated DOK Level 3, in part because of the multiple-choice format.

Standards: Panelists mostly felt that the standards were written at the appropriate level and used this space to comment on the degree that WorkKeys items met or did not meet appropriate specificity or expectations for Grade 12.

Alignment: Panel 1 results indicate 29% felt there was acceptable alignment and 71% felt there needs to be major improvement. Panel 2 results indicate that 25% felt there was acceptable alignment, 25% felt there needs to be slight improvement, and 50% felt there needs to be major improvement. However, as one panelist pointed out, "WorkKeys and NAEP are two different assessments with different purposes. The alignment is acceptable, given the purpose of WorkKeys. I did not expect perfect alignment."

Evaluation of Sub-Study 3 — NAEP to WorkKeys

Coverage of the standards by the items: Panelists' opinions were fairly uniform and blunt about the fact that the NAEP items were not well aligned with the WorkKeys standards. As one panelist commented, "These [WorkKeys] standards and [NAEP] tests are completely mismatched, in item types, content, and focus." Many of the items were conceptual, while the standards were more focused on application and skills. Many items fell outside the scope of the WorkKeys standards, especially items measuring geometry and algebra skills.

DOK levels: In general, panelists mentioned that, considering the number of NAEP items that were codable to WorkKeys objectives (60 for Panel 1 and 89 for Panel 2), the DOK seemed appropriate. For many of the NAEP items that aligned to a WorkKeys objective, the item DOK level matched the DOK level of the objective.

Standards: Many panelists felt that the standards were very specific — possibly too specific — but a few panelists felt they were too general or vague. Some felt that the WorkKeys standards were not at the twelfth-grade level but were instead below this grade level. One panelist noted, "WorkKeys standards are not written for a grade level, they are written with the development of the specific test in mind, and they are written with a "graduated" complexity level goal. So, I think they are appropriate for the purpose for which they were written, but not relevant for NAEP."

Alignment: Panel 1 results indicate 14% felt the alignment needs sight improvement, 71% felt there needs to be major improvement, and 14% felt they were not aligned in any way. Panel 2 results indicate that 38% felt there needs to be major improvement, and 62% felt the two are not aligned in any way. The general opinion of the two panels may be summarized in this comment from one Panel 1 participant: "WorkKeys and NAEP were developed for different audiences and for different purposes and that is apparent in the standards and the assessments."

Evaluation of Sub-Study 4 —WorkKeys to WorkKeys

Coverage of the standards by the items: Most panelists felt that many of the standards were not covered by the items and that some standards were assessed by numerous items. Another idea that was discussed by the panelists is represented in this participant's note: "Many standards did not seem to be covered. On the other hand, many items required using ideas from a lot of the standards, but since I coded using the primary idea behind the item, it appears standards were not covered. In essence, a lot of the standards were used embedded within the ideas contained within the items."

DOK levels: Most panelists indicated that the WorkKeys items were almost entirely at DOK Level 2.

Standards: Panelists felt the standards varied in specificity and had difficulty targeting items to them. Some assumed this was due to their being written for workplace skills rather than for a grade level.

Alignment: Panel 1 results indicate that 100% felt there needs to be major improvement to the alignment. Panel 2 results indicate that 50% felt there was acceptable alignment and 50% felt there needs to be slight 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 Framework has a much broader course of standards, much more like a school's curriculum across high school grades. WorkKeys items are often complex but as pure applied math, and so rarely map to NAEP, or when they do, it is very narrow and really a mismatch of intent. I like NAEP's use of ER and ECR items, which also makes it easier to match its framework."
- "WorkKeys assessment has a different purpose than NAEP. Item types for WorkKeys were more application based and real world, not so much context/abstract/higher thinking. Content coverage was lower on WorkKeys — basic — missing abstract functions/formulas of geometry and algebra. Complexity of WorkKeys was minimal."

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?

- "Students who perform well on NAEP would likely be those who have studied algebra 2 or higher level. Those with more modest scores may have been exposed to algebra 2 but may not have fully developed the concepts. Those who perform poorly probably either were not exposed to very much math at the algebra 2 level or did not understand much of it. I do think that NAEP would give a measure of proficiency for students for each standard. Most students in the US should have taken algebra 2 by 12th grade, so NAEP should be able to differentiate them in terms of achievement. Students who do well on WorkKeys might be those who have been exposed to practical and applied math from living/working on farms or in families including contractors. I do think, however, that some students can do very well on NAEP and not well on WorkKeys. Some topics in WorkKeys are given little or no attention in high school, such as calculating acreage, and relatively little emphasis is placed on unit conversion in high schools."
- "Students who perform well on WorkKeys can do word problems, most of which (but not all) are relatively routine. They may not have many of the high school math skills assessed by NAEP. WorkKeys items are more contextual — we suspect but don't know for sure that students who do well on NAEP would do well on the WorkKeys items."

3) What similarities were identified between the two assessments?

- "Both assessments contained items representing a variety of complexities, with some being extremely simple — perhaps involving only a simple calculation with one operation — while other presented multi-step problems in complicated situations. Both assessments contained a fair number of applied problems as opposed to plain calculations without context. Both assessments use combinations of items with diagrams and other visual representation and items with just verbal descriptions."

4) What differences were identified between the two assessments?

- "The NAEP assessments focused more on geometry, data, and algebra. The WorkKeys focused more on number and measurement. The DOK levels of NAEP ranged from 1 – 3, with the majority Level 2. The WorkKeys had few DOK 3s if any. The DOK 2 levels were "higher 2s" in NAEP. The types of numbers and expression used in the NAEP assessments were more complex. Students were not asked to explain their thinking on the WorkKeys items. Interpreting graphs and data was not often seen in WorkKeys items. The level of mathematical sophistication was higher in the NAEP items."
- "WorkKeys items are all multiple choice, while there are some short-response and constructed-response items on the NAEP. The open-ended questions lend themselves better to DOK Levels 3 and 4. Calculators are allowed on all WorkKeys items, while the NAEP only use calculators on some items. All items on WorkKeys are application problems, while only some NAEP items are presented in real-world application settings."

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

- "It makes sense to use NAEP framework to code WorkKeys items because the skills necessary to take WorkKeys successfully are mostly a subset of the skills necessary to take NAEP. However, it seems difficult to use WorkKeys specifications to code NAEP as NAEP is not a subset of WorkKeys. On the other hand, we shouldn't expect NAEP to code nicely to WorkKeys, so that the process of attempting to do so would confirm it."
- "I remain unconvinced of the value in attempting to compare the 2 assessments given their different purposes, intended audiences and content focuses."

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. As noted in the earlier section on decision rules, one result of having to complete Sub-Study 4 remotely was that only one panelist returned this particular survey, so the sample responses below are all from one participant.

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 assessment has a different purpose than NAEP. Item types for WorkKeys were more application based and real world, not so much context/abstract/higher thinking. Content coverage was lower on WorkKeys — basic — missing abstract functions/formulas of geometry and algebra. Complexity of WorkKeys was minimal." [Identical to response to NAEP final mapping debrief survey]

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?

- "Well-performing WorkKeys students would expect to score poorly → modestly on NAEP assessment…vs. NAEP well performing-modestly [performing] students' performance would demonstrate well on WorkKeys."

3) What similarities were identified between the two assessments?
   - "Some similarities on real world/authentic questions. Very different levels for different audiences."

4) What differences were identified between the two assessments?
   - "WorkKeys is more procedural; NAEP is more conceptual."

5) Please provide any feedback on the usability of the NAEP framework and WorkKeys specifications documents for this alignment task.
   - "The gap is wide — a lot of conceptual knowledge (algebra/geometry) missing from WorkKeys."

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?	3.50
2. How comfortable did you feel with the process of assigning the DOK levels?	3.63
3. How well do you feel Monday's training prepared you for the consensus process?	3.88
4. Overall, how well did Monday's training prepare you for the Alignment (coding) process?	3.50
5. How useful was information about the study you received prior to this week?	2.88
6. How useful were the training and coding materials you received this week?	3.50
7. How qualified did you feel your panel was to conduct this type of alignment?	3.75

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. Perhaps some representatives who were more familiar with WorkKeys might have been helpful.
- The facilitators were effective adjudicators.
- The alignment criteria were moderately useful.
- The WAT software was 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 was 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 *Mathematics* Assessment	WorkKeys *Applied Mathematics* Assessment
Item pool	All 178 items of the 2009 NAEP Grade 12 *Mathematics* item pool were used for this study.	A pool of 58 items drawn from the operational WorkKeys *Applied Mathematics* item pool of hundreds of items was used for this study.
Item context	Items include both pure math and real-world content	All items involve real-world application of math content in a workplace context
Types of items/Average DOK level	• 61% multiple choice / 1.85 • 29% constructed response / 2.14 • 11% multiple part / 1.95 NOTE: Percentages do not equal 100% due to rounding.	• 100% multiple choice / 1.97
Standards on which items are based / Average DOK level	1) Number properties and operations / 1.80 2) Measurement / 1.94 3) Geometry / 2.00 4) Data analysis, statistics, and probability / 2.16 5) Algebra / 2.00	3) A single type of basic mathematics operation; no reordering or extraneous information / 1.20 4) Multiple types of mathematical operations; reordering, extraneous information / 1.29 5) Application of logic and calculation; conversions / 1.29 6) Complex and multiple-step calculations; manipulating formulas / 1.78 7) Nonlinear functions, complex calculations and conversions / 2.14

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 *Mathematics* assessment and the WorkKeys *Applied Mathematics* assessment. Even in the case of Sub-Study 3, where the two panels differed in the number of items unanimously deemed uncodable, the differences were due to the judgment of just one or two panelists for each item, and rater agreement remained high for the sub-study. 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 DOK levels of the two assessments' standards and items:

- The range of DOK levels assigned to the NAEP standards was 1 – 3, and the average DOK level of the NAEP standards was 2.00.
- The range of DOK levels assigned to the WorkKeys standards was 1 – 3, and the average DOK level of the WorkKeys standards was 1.58.
- The range of DOK levels assigned to the NAEP items was 1 – 3, and the average DOK level for all NAEP items was 1.90.

- The range of DOK levels assigned to the WorkKeys items was 1 – 2, and the average DOK level for all WorkKeys items was 1.97.
- The difference between the average NAEP standard DOK level and the average NAEP item DOK level was 0.10, with the standards being at a higher average DOK level.
- The difference between the average WorkKeys standard DOK level and the average WorkKeys item DOK level was 0.39, with the items being at a higher average DOK level.

Key factors likely affecting the DOK levels are related to the fundamental characteristics of the two assessments, some of which are shown in Table 22. The NAEP test is constructed around twelfth-grade math content and application and includes both multiple-choice and constructed-response items, whereas the WorkKeys test is constructed entirely of multiple-choice items around the application of foundational math skills in workplace contexts.

Across the four sub-studies, the NAEP and WorkKeys test items were analyzed for their alignment with the five NAEP standards and the five WorkKeys standards according to four alignment criteria. This produced 80 points for which the degree of alignment was evaluated, using labels of Yes (strong alignment), Weak, and No (not aligned).

For 78 of these 80 points, the two panels agreed on the degree of alignment. For one of the two points of difference — in Sub-Study 2, for Standard 1, the Balance of Representation criterion — one panel's assessment was that there was weak alignment while the outcome of the other panel was no alignment. The index values are very close here, however, just on either side of the threshold value for weak alignment. And for the remaining point — Sub-Study 2, Standard 3, Depth-of-Knowledge Consistency — the panels came to opposite conclusions (Yes vs. No). This difference, however, is somewhat of a statistical artifact due to the coding of one item by one rater. Thus, the agreement between the two panels is very high overall.

After the conclusion of the panel meetings, per Dr. Webb's alignment methodology, 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. After the facilitators' thorough efforts, there remained only the two points 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 found that there is alignment for all five standards using all four alignment criteria. That is, there were six or more items that targeted each of the five standards, the majority of those items were at or above the DOK levels of the objectives to which they were coded, at least 50% of all objectives within each standard were hit by at least one item, and the number of hits was fairly evenly distributed across the objectives hit. There were no items the panelists deemed uncodable.

The NAEP objectives emphasized most by the NAEP items are those related to solving application problems involving numbers, applying geometric properties and relationships to solve problems, reading or interpreting data in graphical or tabular format, and algebra. The objectives targeted least by the items were spread across all of the standards except Standard 5, Algebra.

Sub-Study 2, the alignment of the WorkKeys items to the NAEP standards
Sub-study 2 found some degree of alignment of WorkKeys items to NAEP Standard 1, “Number Properties and Operations”; and to Standard 2, “Measurement.” The data indicate weak alignment for Standard 4, “Data Analysis, Statistics, and Probability,” and no alignment for Standard 3, “Geometry,” or Standard 5, “Algebra.” No items were coded to Standard 5, “Algebra,” and only one panelist coded one item to an objective within Standard 3, “Geometry.” There were two WorkKeys items coded to generic NAEP standards, and no items that were deemed uncodable.

The NAEP objectives targeted by the most WorkKeys items include problem-solving applications of number operations and measurement. The NAEP objectives to which no WorkKeys items aligned are primarily related to geometry; data analysis, statistics, and probability; and algebra.

Sub-Study 3, the alignment of the NAEP items to the WorkKeys standards
Sub-Study 3 found alignment or weak alignment at 75% of the 40 points for which the alignment was calculated; no alignment was found at the remaining 25% of the points. Fifty percent of the NAEP items were deemed uncodable to the WorkKeys standards by all panelists.

Findings of this sub-study include that the WorkKeys objectives most frequently covered by NAEP items include geometry content; fractions, ratios, percentages, or mixed numbers; and basic statistical concepts. The WorkKeys objectives that were targeted least often by the NAEP items range across a variety of workplace-oriented math skills: conversions, determining the best deal, finding errors, and calculating discounts or markups.

Sub-Study 4, the alignment of the WorkKeys items to the WorkKeys standards
Sub-study 4 found alignment at 80% of the points at which alignment was calculated, weak alignment at 5% of the points, and no alignment at 15% of the points. 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 58 items — accounting for two items in common between the two test forms used for this study — from the entire operational WorkKeys pool of hundreds of items), rather than a more extensive item pool. This decreased the likelihood that the sampled items would completely cover all WorkKeys objectives. As a point of contrast, 178 NAEP items were used in the study.

The WorkKeys objectives targeted most frequently include those involving one or two operations; ratios, percentages, or mixed numbers; calculating areas along with unit conversions; and determining the best deal. The objectives targeted least often by the WorkKeys items are those involving negative numbers, finding mistakes, the volume of objects, nonlinear functions, and basic statistical concepts.

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.

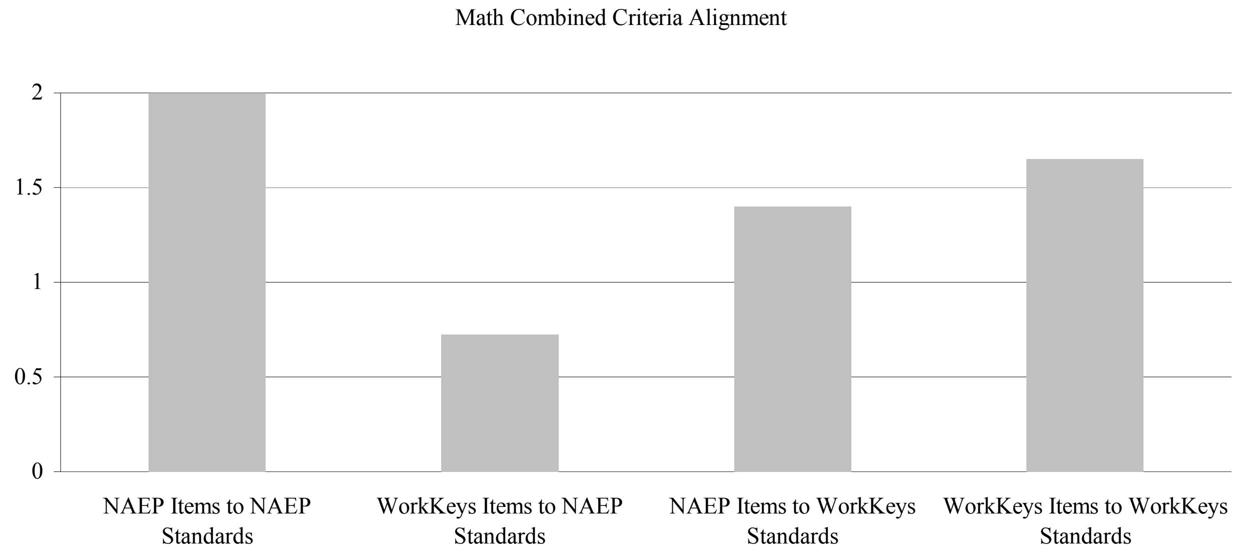
Table 23: Content distribution summary*

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) Number properties and operations	14%	13%	66%	64%
2) Measurement	11%	12%	28%	29%
3) Geometry	18%	19%	0%	0%
4) Data analysis, statistics, and probability	24%	24%	5%	7%
5) Algebra	33%	33%	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) A single type of basic mathematics operation; no reordering or extraneous information	1%	0%	17%	17%
4) Multiple types of mathematical operations; reordering, extraneous information	22%	23%	33%	32%
5) Application of logic and calculation; conversions	19%	18%	20%	19%
6) Complex and multiple-step calculations; manipulating formulas	18%	17%	27%	28%
7) Nonlinear functions, complex calculations and conversions	40%	41%	3%	3%

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

The following 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	2.00	2.00
2) WorkKeys items to NAEP standards	0.80	1.40	0.00	0.70	0.72
3) NAEP items to WorkKeys standards	1.60	2.00	0.80	1.20	1.40
4) WorkKeys items to WorkKeys standards	1.60	2.00	1.00	2.00	1.65

General conclusions

The results of the four sub-studies and the blueprint analyses may be summarized by saying that the content represented in the WorkKeys standards and test items used for this study comprise a subset of the content represented in the NAEP standards, since all of the WorkKeys items were codable to NAEP objectives. However, this description is imperfect, because the results of this study also show that not all of the WorkKeys standards are covered by the NAEP assessment. In other words, taken together, the blueprint analyses and four sub-studies show that not all of the

WorkKeys content is included in the scope of the NAEP standards. There are WorkKeys objectives that are not included in the NAEP framework.

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

A) The NAEP Grade 12 *Mathematics* assessment covers a broader range of math skills than does the WorkKeys *Applied Mathematics* assessment, particularly in geometry; data analysis, statistics, and probability; and algebra.

B) The WorkKeys *Applied Mathematics* assessment focuses on a narrower range of math skills than does the NAEP assessment. Specifically, the WorkKeys assessment focuses on the application of foundational math skills in workplace situations.

C) Most of the WorkKeys items aligned with NAEP objectives related to number operations and measurement.

D) WorkKeys objectives that are not assessed by the NAEP items include conversions, determining the best deal, finding errors, and calculating discounts or markups. Thus, even though the math content included in the NAEP framework is significantly broader than that included in the WorkKeys framework, there are numerous WorkKeys objectives that the NAEP items used in this study did not target.

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.

The representation of the NAEP framework used for this study was very challenging. With 154 separate objectives to choose from for coding items, the coding task was difficult and grueling, and the results were at a very fine-grained level. If additional alignment studies are conducted with the NAEP assessment in the future, ACT would recommend that consideration be given to how the NAEP framework might be represented with fewer objectives. Perhaps objectives could be grouped together to create slightly broader categories, reducing the number of options panelists must consider when coding items and also the amount of time required to complete the task.

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 strong, weak, and not aligned points in the standards becomes less clear.

This potential for misleading results underscores the necessity to consider the four alignment criteria in concert. That is, no single criterion of the four provides a complete picture of the alignment, and it is necessary to consider all four together to achieve a more accurate understanding of the degree of alignment. Another case in point is the Depth-of-Knowledge Consistency result for Sub-Study 2. In this instance, Panel 1 results show 100% consistency for Standard 3, while Panel 2 results show 0% consistency for Standard 3. A closer examination of the data reveals that this result is due to one Panel 1 member coding one item to one objective within Standard 3, while no other panelist in either group coded any items to any other objectives within the standard. Because the one item was coded with the same DOK level as the one objective, the consistency was 100%. Considering this result in the context of the other alignment criteria results, however, shows that the overall results of the two panels are much more closely aligned than this one result would indicate.

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. The current methodology relies primarily on identifying the objectives and standards to which the most items have been coded, and this may indeed be the best primary criterion to use in determining alignment. However, it would likely be useful to future researchers engaged in this type of study if there were other established markers and threshold values to use as criteria for judging the degree of alignment between two assessments. 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. (January, 2010) *Grade 12 NAEP math assessment and WorkKeys Applied Mathematics assessment blueprint analysis*. Iowa City, IA: Author.

ACT. (2008). *Applied Mathematics technical manual.* Iowa City, IA: Author.

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). *Mathematics 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.

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