Design of Content Alignment Studies in Mathematics and Reading for 12th Grade NAEP and other Assessments to be used in Preparedness Research Studies[1]

The purpose of this paper is to provide a detailed design to guide implementation of content alignment studies for the grade 12 National Assessment of Educational Progress (NAEP) in reading and mathematics with respect to other assessments that the National Assessment Governing Board plans to use to provide indicators for reporting preparedness of 12th graders on NAEP in these subjects. The alignment studies are to form a part of the evidence in a series of research studies designed to explore NAEP's capacity to produce and report valid data on the preparedness of 12th graders for post-secondary activities.

This design document addresses all key points that must be considered for implementing a content alignment study between two tests. NAEP is a highly visible assessment program, and the alignment studies are central to the 12th grade preparedness research. Because different assessments will be used, the Governing Board faces the challenge of developing alignment studies that produce comparable information. The Board wants to generate as much information as possible about the content relationship and alignment between NAEP and the other assessments of interest while also assuring that the information comparing NAEP across assessments is comparable. Whatever the process used to judge the content alignment between two assessments, the process should be transparent and replicable.

[1] The National Assessment Governing Board contracted the services of Norman L. Webb, Senior Research Scientist Emeritus, Wisconsin Center for Education Research, University of Wisconsin-Madison to develop this design document for use in a series of content alignment studies for the Grade 12 National Assessment of Educational Progress in reading and mathematics. Mr. Webb delivered a complete draft to the Governing Board on December 19, 2008. The draft document was reviewed extensively in January and February, 2009; and the Governing Board approved the design for implementation in the content alignment studies at the March 2009 meeting. Several modifications were made by Governing Board Staff to clarify specific points, to more fully reflect the Board's goals for the studies, and to respond to recommendations from reviewers. Staff thanks Mr. Webb for his generous assistance throughout this process and the many reviewers who helped the Board to reach closure on the choice of methodologies.

Alignment

Alignment in the current context of No Child Left Behind 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).

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, 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.

Methods for Conducting Alignment Studies

Three methods represent the most prevalent approaches for judging the alignment between assessments and standards (Le Marca, Redfield, Winter, & Despriet, 2000). All three approaches employ from five to eight content experts as the panelists whose alignment judgments are used to determine the degree of alignment. One way that the approaches differ is in the judgments made by the panelists. In the process developed by Webb (2002), panelists assign the depth-of-knowledge level (level of complexity) to each objective underlying each content standard. Next the panelists map each item to the standards. The two steps in mapping items to content statements include having panelists independently assign a DOK level to an item on the assessment and then assign the item to up to three objectives. Panelists are to map an item to an objective only if content knowledge expected to satisfy the objective is necessary, at least in part, to answer the item correctly.

The Survey of the Enacted Curriculum (SEC) process, developed by Porter and colleagues, uses a comprehensive matrix of content topics by cognitive levels to analyze the content from different documents using a common content language (Porter, 2002 & 2006). Panelists map each objective underlying the standards to the cell in the matrix representing the most appropriate topic and cognitive level. Panelists can assign one objective to more than one cell as appropriate. Panelists also map each item to the appropriate topic-by-cognitive-level cell. The alignment, reported as an index value between 0 and 1, is the aggregation of the proportion of cells in common between the mapping of the content standards, the assessment, and/or the teacher’s instructional objectives to the matrix. In this way it is possible to compare standards with assessments, and each of these with the enacted curriculum as described by the teacher.

The Achieve, Inc. protocol for analyzing the alignment between an assessment and content standards uses panelists to produce information on four alignment criteria—content centrality,

performance centrality, range and balance, and challenge (Rothman, Slattery, Vranek, & Resnick, 2002). The analysis begins with a content expert verifying the state's own alignment between the assessment and standards such as would be described in a test blueprint. Then panelists analyze and reach consensus on the relationship between each item and its assigned standard and objective as specified by the blueprint. Panelists reach consensus on the four alignment criteria for each item (content centrality, performance centrality, range and balance, and challenge). For content and performance centrality, panelists can agree that the item fully addresses the intent of the assigned objective, partially addresses the intent, or in no way addresses students' knowledge as expressed by the objective. Results are reported as the percentage of items with full, partial, or no content and performance centrality; whether the collection of the items is appropriately challenging to students at the given grade level; and whether some topics are over- or under- represented.

The three alignment procedures vary in terms of the information produced on the relationship between assessments and standards. The findings of the relationship between the assessment and standards from the Webb process are reported using four alignment criteria—Categorical Concurrence, Depth-of-Knowledge Consistency, Range-of-Knowledge Correspondence, and Balance of Representation. A distinguishing factor of the Webb process is that specific decision rules are used to determine if the alignment between content standards and the assessment is acceptable. The SEC process produces an index, ranging in value between 0 and 1, representing the overall alignment between the standards, the assessment, and the classroom curriculum. Findings from SEC are also reported as topographical maps and in other data displays, one for each document (standards, assessment, and curriculum) analyzed. The topographic maps can be viewed side-by-side to determine the variation in emphasis from classroom to assessment program to standards of topics by cognitive levels. Results derived from the Achieve, Inc. protocol are reported in a narrative including some tables showing results for the alignment attributes. The narrative reports the degree of alignment as determined through the consensus process, how the alignment could be improved, and any other relevant information.

Any of these three methods could be used to analyze the alignment between two assessments. However, the purpose for analyzing the NAEP with assessments of post-secondary education preparedness is to provide supporting information for the valid use of other assessments with grade 12 NAEP to interpret results and report findings regarding students' preparedness for higher education and workplace training. The Webb process, the most popular approach among states for comparing standards and assessments (Porter, 2006), provides independent judgments among panelists on the degree of alignment using multiple criteria—topic, complexity, range, and balance. The assessments can be mapped directly to the NAEP assessment framework. This produces information on the content within an objective or subtopic that is or is not targeted and it uses the terminology of the actual framework. Mapping both assessments to the same framework (e.g mapping both the NAEP and SAT mathematics assessments to the NAEP mathematics framework), the assessments can be compared according to the number of assessment items mapped to each content area, subtopic, and more detailed content levels; distribution of items from each assessment within each content area by levels of complexity; proportion of subtopics with at least some items from each assessment; and balance in emphasis (over or under) by assessment items of any objectives under a subtopic and content area in relationship to other objectives.

Both SEC and Webb system panelists independently analyze the assessments; but rather than mapping the items directly to an assessment framework, as in the Webb system, SEC panelists map items to a common framework or "language system." An advantage of the SEC is that an assessment would only need to be mapped to the SEC content-by-cognitive level framework. With the Webb process, an assessment would have to be mapped to each framework used in the comparison. After an assessment has been mapped with the SEC method, the assessment could be compared to any other document (assessment, framework, or curriculum) that has also been mapped to the SEC framework. A disadvantage to this method is that the alignment between documents is reported as a single index describing a holistic relationship between documents. However, graphic representations of the mappings can be displayed to represent comparability of specific topics by cognitive levels for any of the alignments examined.

The Achieve system depends heavily on verifying the alignment of the assessment to a blueprint or framework, and the protocol would require major modifications to be adapted for an assessment to assessment analysis. The Achieve methodology would be less suitable for the goals of the Governing Board than either of the other two common alignment procedures.

Both the Webb process and the SEC have advantages and disadvantages. Both have computerized tools that can be used to enter and analyze data. Both produce measures of reliability among panelists. The SEC would require fewer analyses, but would produce less information on the degree of alignment. The Webb process will require mapping assessments to different frameworks, but will produce more detailed information. Both methods are transparent and replicable. Of these two systems, the Webb process is more suited to the Governing Board's goal of maximizing information about the degree to which the NAEP assessments are aligned with other assessments.

Alignment of NAEP Assessments to Other Assessments

Different methods can be used for judging the alignment of the NAEP assessments in reading and mathematics with assessments measuring preparedness for post-secondary activities. The Webb process is a content analysis. Two assessments are aligned to the degree that the two assessments are judged by a group of panelists to target the same content knowledge at a similar level of complexity. Note that content complexity is different from content difficulty. Content complexity is influenced by the structure of the content and performance expectations. An assessment item is more complex if the item requires knowledge of multiple concepts and ideas, if the answer can be derived in many ways, and if generalization is required. Difficulty is a psychometric term related to student performance on an item and is reported as the percentage of students who correctly answer an item. Difficulty is related to complexity, but it can depend on other factors such as the speediness of the test, opportunity to learn, and item format.

Most tests of student content knowledge are composed of a sample of items from some content domain. It is possible to have distinct tests that serve common purposes and produce comparable measures of students' content knowledge. For two or more tests to have content alignment and

similar content coverage, the tests should sample content knowledge from the same content domain.

Alignment criteria (Webb, 1997) used to analyze the alignment between tests and curriculum documents can also be used to judge the alignment between two or more tests:

- Categorical Concurrence—The same or consistent categories of content appear in both assessments.
- Depth-of-Knowledge Consistency—The same depth of content knowledge is elicited from students by both assessments.
- Range-of-Knowledge Correspondence—A comparable span of knowledge within topics and categories is targeted by both assessments.
- Balance of Representation—A similar emphasis, indicated by the number and weighting of assessment items, is given to different content topics and subtopics on each assessment.

In judging the alignment between two assessments, these alignment criteria should be applied relative to a content domain. A test of students' content knowledge generally is designed to produce information on student performance related to a content domain by sampling content knowledge. The results from the assessment are used to make inferences about student knowledge relative to a content domain, as generally described in an assessment framework or blueprint. Because of the vastness of the possible items that could be used to assess students' knowledge of a domain, it is unlikely that any two assessments targeting the same domain will have precisely the same items. Thus, any item-by-item comparison between two assessments could result in a minimal match between the assessments. The likelihood of an item-by-item match between two assessments would be expected to decrease as the differences in the purposes of the two assessments increase. NAEP is designed to monitor educational progress in the nation, whereas other tests of interest to 12th grade NAEP preparedness research are designed with a more narrow purpose of predicting success of students in higher education or placing students in college courses, for example.

The approach for analyzing the alignment of the NAEP mathematics and reading assessments to other assessments, as described here, is designed to compare the assessments by how the items represent content domains. For mathematics, five content areas specified in the 2009 NAEP Mathematics Framework (National Assessment Governing Board, 2008a) serve well as content domains for comparing the alignment between two or more tests:

1. Number Properties and Operations
2. Measurement
3. Geometry
4. Data Analysis, Statistics, and Probability
5. Algebra

Exhibit 1: Content areas specified in the 2009 NAEP Mathematics Framework.

For reading, the cross-section of the aspects of reading and the context of reading specified in the NAEP Reading Framework for 2009 (National Assessment Governing Board, 2008b) can serve as content domains. The aspects of reading in the Reading Framework are:

1. Locate and recall
2. Integrate and interpret
3. Critique and evaluate

The text types are represented in the text matrix below.

Grade 12 Reading Text Matrix

	Genre/Type of Text	Text Structures and Features	Author's Craft
Fiction	• Satire • Parody • Allegory • Monologue Plus increasingly complex application of grades 4 and 8	Organization • Differentiation of plot structures for different purposes and audiences Elements • Interior monologue • Unreliable narrators • Multiple points of view Plus increasingly complex application of grades 4 and 8	• Dramatic irony • Character foils • Comic relief • Unconventional use of language Plus increasingly complex application of grades 4 and 8
Literary Non-Fiction	• Classical essay Plus increasingly complex application of grades 4 and 8	Increasingly complex application of grade 4	• Denotation • Connotation Plus increasingly complex application of grades 4 and 8
Poetry	• Sonnet • Elegy Plus increasingly complex application of grades 4 and 8	Elements • Complex themes • Multiple points of view • Interior monologue • Soliloquy • Iambic pentameter Plus increasingly complex application of grades 4 and 8	• Denotation • Connotation • Irony • Tone • Complex symbolism • Extended metaphor and analogy Plus increasingly complex application of grades 4 and 8
Exposition	• Essay (e.g., political, social, historical, scientific, natural history) • Literary analysis Plus increasingly complex application of grades 4 and 8	Increasingly complex application of grade 4	• Denotation • Connotation • Complex symbolism • Extended metaphor and analogy • Paradox • Contradictions and incongruities • Ambiguity Increasingly complex application of grades 4 and 8

Exhibit 2: Type of Text, Text Structures and Features, and Author's Craft for Grade 12 from the 2009 NAEP Reading Framework.

For mathematics, the content areas are further delineated by subtopics and objectives. These more precise statements of content knowledge can then be used to compare the range or span of knowledge within content areas assessed by a test. For reading, the element of texts, reading skills, and reading passages add more detailed specifications, although neither the framework nor specifications provides detailed objectives for student achievement. The preliminary achievement levels definitions for reading are a potential source of this level of detail for reading.

The process for analyzing the alignment between NAEP and other assessments is designed to determine the *degree* of alignment. Most likely, two assessments will overlap in content coverage with some content common to both assessments and other content unique to each assessment (Exhibit 3).

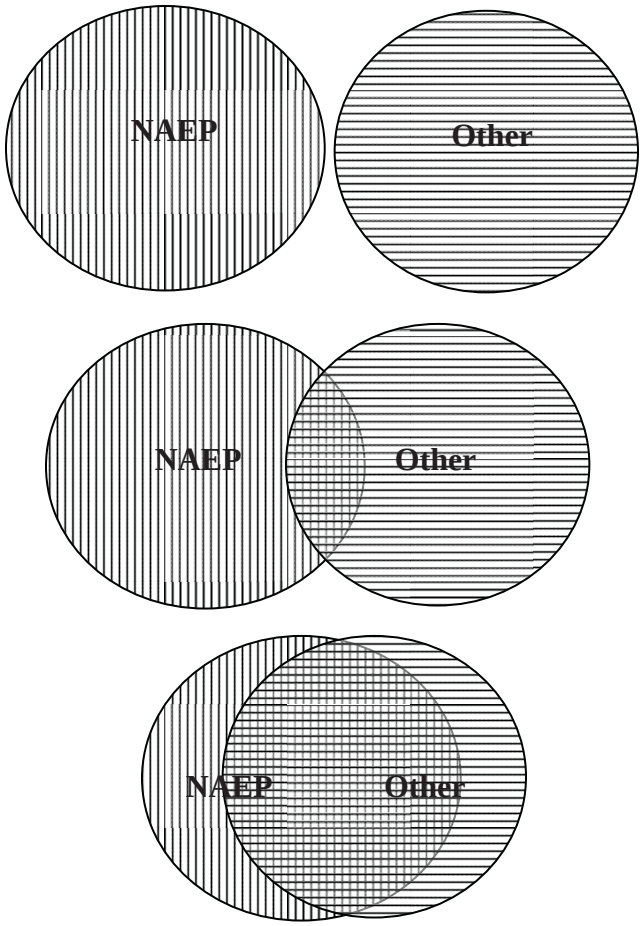

Exhibit 3: Depiction of different degrees of alignment between NAEP assessment and another assessment.

The purpose of the alignment analysis is to determine both the extent of the overlapping content knowledge targeted by each assessments and the extent of the content knowledge that is unique to each assessment. The alignment criteria provide a basis for reporting what is common between

two assessments and what is different—the categories or topics, the depth-of-knowledge or cognitive level, the range or breadth, and the degree of emphasis. The process includes using the NAEP framework as a representation of the content along with using the framework of the other assessment as a representation of the content. This will allow each assessment to be compared using the language system of the NAEP framework and the language system of the framework for the other assessment referred to hereafter as *Pexam*. This bidirectional analysis will be particularly helpful in determining the categorical concurrence, range, and balance of each assessment relative to each framework and to each other. It is possible that the analysis will show there is little or no alignment between the NAEP assessments and any Pexam.

Determining the Degree of Alignment Using the Four Criteria

The categorical-concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. *The criterion of categorical concurrence between assessments is met if the same or consistent categories of content appear in both assessments.* This criterion is judged by determining the number of items each assessment includes for each content area and subtopic. Two assessments agree in categorical-concurrence if the proportion of items from each assessment assigned to each content category is similar.

Two assessments can be aligned not only on the basis of the content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge consistency between two assessments indicates alignment if the cognitive demand of the two assessments is approximately equal.* For consistency to exist between two assessments, as judged in this analysis, the proportion of items at each level of complexity should be similar for the main content categories and subcategories.

For two assessments to be aligned, the breadth of knowledge required on both should be the same, or very nearly so. The range-of-knowledge criterion is used to judge whether a span of knowledge expected of students on one assessment is the same as, or very nearly the same as, the span of knowledge expected of students on the other assessment. The range criterion considers the proportion of subcategories (e.g. subtopics or objectives) under a content category (e.g. content area or standard) with at least one corresponding assessment item. The range of knowledge is comparable between two assessments if the proportion of subtopics assessed is the same or similar.

In addition to comparable depth and breadth of knowledge, aligned assessments require that knowledge be distributed equally in both. The range-of-knowledge criterion only considers the number of subcategories within a content category hit (a subtopic with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among the subcategories (e.g. subtopics or objectives). *The balance-of-representation criterion is used to indicate the degree to which one content subcategory is given more emphasis on one assessment than the other assessment.* An index is used to judge the distribution of assessment items among subcategories underlying a content category. 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. 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. Two assessments have comparable balance of representation if the distribution of items among subcategories is the same as determined by a comparable index value.

The overall alignment between two assessments is determined by similar values on all four alignment criteria.

Design of Alignment Study

The major components of an alignment study to be addressed in this design include:

- specification of the content domains for the comparison of two assessments
- specification of the criteria to be used to determine the degree of alignment between two assessments
- process for panelists to conduct the analysis
- means for analyzing and reporting the findings

The Webb alignment process will be used for analyzing the alignment between the 12th grade NAEP in mathematics and reading and the post-secondary assessments (Pexams). This process includes using the depth-of-knowledge (DOK) levels definitions (see Appendices B and C) in mathematics and reading to assign levels of complexity to assessment items and objectives; having a group of trained panel members conduct the analysis; assigning levels of complexity to objectives or expectations in each assessment framework; assigning DOK levels and content objectives to assessment items; and analyzing and reporting the results using four alignment criteria (categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation). The Web Alignment Tool (WAT) (http://wat.wceruw.org/) is recommended for collecting the data from panelists and conducting analyses, but the choice of analysis instrument is not critical to the outcome of the study so long as all the data are collected and computations are performed in the same manner.

A key reason for analyzing the alignment between two assessments is to determine the extent to which the two assessments target the same content domains or the extent to which inferences can be drawn from students' performance on one assessment regarding their capacity to perform in a comparable content domain on another assessment. The NAEP frameworks for both mathematics and reading specify the content domains to be used in developing items and selecting them for the NAEP. Using these NAEP frameworks as the content structure provides one means of analyzing alignment and drawing conclusions about alignment between the NAEP assessments and other assessments mapped to the NAEP framework. Using the framework of the other assessments to be compared to NAEP provides another basis for the analysis of the relationship between the two assessments.

In the Webb alignment approach, panelists map items from an assessment directly to the assessment framework. After the items are mapped to the framework, it is then possible to describe the content categories (objectives, topics, and so forth) that were not targeted by the assessment. The purpose of mapping items to the assessment framework is not to evaluate the quality of the items or the validity of the assessment; rather, it is to establish a basis for

comparing the two assessments. This process of mapping items to frameworks results in additional information on the match of an assessment to a framework that is more detailed than would be the case if the alignment were based on framework objectives or other higher-level attributes of the framework or item specifications. Four item mapping/coding procedures for each subject assessment are called for in this design:

1. NAEP items to the NAEP framework
2. Pexam items to the NAEP framework
3. Pexam items to the Pexam framework
4. NAEP items to the Pexam framework

Panels

The study is to have two groups of six to eight panel members each for each subject independently analyze the assessment frameworks and the assessment items during a period of approximately five days. The panels should be equivalent in terms of area of content expertise, level of content expertise (secondary/post-secondary), and demographic attributes. Racial-ethnic and geographic diversity should characterize the panels.

Data are to be analyzed to determine the consistency in the results for the two groups. The two groups will initially and primarily operate independently. The results from the two groups will serve as a replication of the alignment judgments.

Having two groups complete the alignment analysis concurrently allows a real-time check on the replicability (i.e., the reliability) of the findings. If the findings from both groups are comparable, then greater confidence is assigned to the results. Having the groups perform the analysis at the same time allows the opportunity for on-site adjudication and resolution regarding how specific aspects of assessments are to be interpreted. Decision rules must be developed in advance so that instructions can be prepared to train panelists and avoid ambiguous situations that may be confusing and inefficient. In the event that questions arise, however, the alignment results will be based on the on-site resolution and adjudicated data collected for the two panels.

To evaluate the content alignment of 12th grade NAEP to the other assessments, several tasks must be accomplished. The following tasks are to be included in each study. An agenda is included in Appendix A to provide an estimate of the amount of time needed for the various tasks included in this study.

Tasks

Task 1 Date and location for conducting the studies set, including arrangements for required meeting facilities.

Task 2 Qualified panel members (6-8 for each of two replicate panels for each subject) recruited and confirmed; one expert group facilitator for each replicate panel (2 for each subject) contracted.

Task 3 Materials prepared for training panelists and collecting and recording data. Data analysis software (e.g. WAT) prepared by entering the components of each framework into the software which will have been customized to capture findings in Task 4 (comparisons of test specification documents) and to capture findings of partial coverage along with codes for panelists' rationales for alignment judgments.

Task 4	Comparative analysis of the pairs of test blueprints (NAEP and Pexam) conducted by an expert for each subject.
Task 5	Panel members trained in DOK level definitions and assignment of items to key framework components for each assessment for each subject.
Task 6	Panel members trained to use the WAT (or comparable software) features and procedures.
Task 7	Panel members assign DOK levels to NAEP framework components and reach consensus on these. DOK agreement is reached between the DOK levels assigned by the replicate groups.
Task 8	Panel members map 2009 NAEP item pool to the grade 12 NAEP framework objectives for the subject.
Task 9	Panel members respond to de-briefing questionnaire about alignment of NAEP item pool to NAEP framework.
Task 10	Facilitators review codings and determine whether there are discrepancies in assigning items to objectives and in the results on the four alignment criteria; panelists adjudicate discrepancies.
Task 11	Panel members map each of two forms of the Pexam to grade 12 NAEP framework.
Task 12	Panel members respond to de-briefing questionnaire about alignment of Pexam items to NAEP framework
Task 13	Facilitators review codings and determine whether there are discrepancies in assigning items to objectives and in the results on the four alignment criteria; panelists adjudicate discrepancies
Task 14	Panel members complete final debriefing questions about the content similarities and differences between the NAEP items and the Pexam items relative to the NAEP framework.
Task 15 - Task 22	Same as Tasks 7-14, using Pexam framework for mapping items for evaluation of alignment.
Task 23	Alignment study team analyzes the data collected at the study and the document comparisons of the NAEP and Pexam assessment frameworks.
Task 24	Alignment study team writes the final reports indicating how the NAEP and Pexam assessments are aligned and how the two assessments are not aligned. There will be one report for each subject assessment.

Some of the tasks listed above are explained in more detail below.

Task 4 Comparative Analysis of Test Blueprints: An expert for each subject will conduct a comparative analysis of the pairs of test blueprints (NAEP and Pexam). The comparative analysis of the test blueprints for NAEP and the available blueprints for all other tests to be included in the analysis is to be done prior to the item analysis. The main purpose of the blueprint comparative analysis is to identify the similarities and differences in the content specifications, item types, reading passages, and other specifications used in the design of each assessment. The comparative analysis is to specify the content organization for identifying items to be included on the NAEP assessment and the Pexam comparison assessments. For example, the mathematics framework for NAEP organizes the mathematics domain into five content areas which are further divided into subtopics and objectives (National Assessment Governing Board,

2008a). The reading framework for NAEP organizes the reading domain by type of texts (fiction, non-fiction, expository, etc.) and features of texts. Within the cells formed by the types of texts and the features of texts, content is further specified by skills and elements--such as theme, major characters, and major events). (National Assessment Governing Board, 2008b).

In the comparative analysis, a side-by-side chart of the content organization is to be prepared that will display how the content structure used for the construction of each assessment is the same or different. The content comparative analysis is to identify differences in the topics included in one set of specifications but not in the other, such as the range and type of numbers for mathematics and the elements in reading. The analysis is also to indicate the grain size, or degree of specificity, in identifying the content for each assessment and how these are similar or different. It is possible that the content specifications between NAEP and another assessment address the same topics, but that one set of specifications does so at a more sophisticated level of specificity. In addition, the content comparative analysis is to determine how the performance specified in one framework is expected to differ from the performance in the other framework. One framework may specify that students are to be assessed on *writing a variety of numbers*, whereas the other framework specifies that students are to be assessed on *reading and writing numbers*. Finally, the content comparative analysis using a side-by-side chart should point out any inconsistencies found within each of the frameworks included in the analysis. For example, a standard may state that students are to analyze characteristics of real numbers, whereas all of the underlying objectives only require that students represent or use applications involving rational numbers.

The comparative analysis should also identify other characteristics of items as specified in the assessment framework and test specifications documents. The characteristics should include:

1. Number and proportion of items for each item format (multiple choice, short constructed-response, extended constructed-response, and any other types of items)
2. Scoring rubrics and rules for constructed-response items
3. Resources available to students (e.g. calculators, dictionaries, etc.)
4. Reading difficulty and grade-level targeted by items
5. Information about reading passages (original source, authentic texts, length, number of items per passage, organization of items within passages, etc.)
6. Information about test administration (when the assessments are administered, amount of time targeted for the assessments, time constraints, accommodations allowed, and the like)

The comparative analysis should be fully documented and presented in an interim report. (See the *Reports* section on page 25.)

Task 5 Training of Panel Members: Panel members need to be fully trained for the alignment tasks. The training should begin with an overview of the alignment process. The overview should include instruction in the following features of the process:

1. What is meant by alignment between an assessment and an assessment framework and between two assessments
2. The four alignment criteria used to determine the degree of alignment
3. Levels used for each criterion to specify the acceptable alignment overall
4. The steps in the alignment evaluation process

5. The general definition of depth-of-knowledge (DOK) used to identify content complexity, as well as specific definitions for the assessment
6. Illustrations of DOK levels assigned to content expectations (framework level) and specific objectives and items
7. Illustrations of DOK levels assigned to assessment items of each type
8. Coding rules, including the maximum number of content expectations to which one item may be coded, the implications for coding items to more than one content expectation, requirements for coding an item to a specific content expectation
9. How to produce good notes to document coding rationales, questions, and so forth
10. Source-of-challenge issues that should be noted, such as construct irrelevant features that may inadvertently cause an item to be more or less difficult or shift the cognitive demand away from the intended target.
11. The use of generic objectives in the event that a panelist judges that an item does not fit any content objective or expectation
12. Login procedures and navigation guidance for the data entry and analysis software (e.g., the WAT)
13. Other administrative details

The subject matter facilitators should determine when the group has a sufficient understanding of the DOK levels. It is not necessary for all panel members to have a precise understanding of the DOK levels until panelists are about to assign DOK levels to the NAEP objectives and elements/skills in the NAEP framework (or the first DOK assignment task).

Task 6 Use of the Software/Analysis Tool: The panel members should logon to the WAT (http://wat.wceruw.org/) or similar software analysis tool selected for this purpose. The analysis tool must be configured before the alignment panelists are convened so that members of each group are registered in the system and ready for login. Quality control procedures and advance planning are essential for the successful use of this tool by panelists.

Task 7 Assign DOK Levels to NAEP Objectives: Panel members in each of the two replication groups will independently assign DOK levels to the objectives under the content areas and subtopics for mathematics or the elements and skills under the contexts and aspects for reading. Panelists will use the WAT or other software/analysis tool to record the DOK levels assigned to each objective/element/skill in the NAEP framework.

Once all of the panelists have coded the DOK for each objective/element/skill, the group facilitator will print the results, listing the code assigned for each panel member. Any objective/element/skill without full agreement should be discussed to reach consensus for the group on the assigned DOK level. Reaching true consensus among panel members is an important goal because the process affords the panel members the opportunity to discuss the fine points for each objective/element/skill. The group facilitator must be trained in the process and assure that all panel members provide input.

After the two groups have determined the DOK levels for each objective/element/skill listed in the NAEP framework, the two group facilitators will meet to review the results and identify any differences between the two groups. The group facilitators will discuss the rationales provided

in each group and decide on the DOK level with the most compelling reasons. In the absence of a compelling reason, the DOK level assigned by the majority across the two groups will be used. Note that this adjudication process is conducted by the group facilitators and requires time in the agenda when panelists are not convened. The final DOK level assignments will be reported to panel members and panelists have the opportunity for discussion.

Task 8 Map the NAEP Items to the NAEP Assessment Framework: Next, the panel members should independently map the full 2009 NAEP item pool to the NAEP framework. To assure that the panel members are comfortable with this process, the group facilitator should select several items from the assessment for panel members to code for practice (individually, with pencil and paper). The sample items should be selected to represent the range of content, item formats, and other aspects of the assessment. The group facilitator will then have the panel members review and discuss briefly the codes assigned to map these items to assure panelists understand and agree on the procedures for coding items.

Once the facilitator is comfortable that panel members are correctly mapping items to the objectives/elements/skills, the panelists continue mapping the items by assigning a DOK level to each item and mapping the item to up to three objectives. A questionnaire should be developed to document the level of understanding and confidence panelists have before starting the coding process and after they have completed the task. In assigning an item to an objective, it is important that at least some of the content addressed in the objective is *necessary* in order to answer the item correctly. An item should not be mapped to an objective if the content knowledge in the objective is only relevant—and not necessary. For example, a question may require students to interpret the slope between two points although students could correctly answer the question by constructing an equation. This item should be assigned only to an objective about "determining the slope of a line" and not to an objective about "writing a linear equation."

It is critical that panel members apply the rule that content knowledge as expressed in an objective is absolutely necessary to answer the item correctly in order to map an item in the objective. If content knowledge from more than one objective is absolutely necessary to correctly answer an item, then the item can be assigned to one primary objective and up to two secondary objectives. When an item is assigned to multiple objectives, the item is weighted by the number of objectives. That is, the computations for Categorical Concurrence, Range, and Balance will incorporate all of the assigned objectives. If one item is assigned to two objectives, then both objectives are counted as a hit for Categorical Concurrence.

Panel members typically need about two minutes to code a multiple-choice item and about five minutes to code a constructed-response item. For constructed-response items, the panelists should consult the scoring rubric and the anchor items to assign codes.

Task 9 Panel Members Respond to a De-Briefing Questionnaire: Panel members should individually respond to a small set of questions after the items have been coded to the NAEP objectives. These debriefing questions are designed to elicit from the panel members more detailed information about how the collection of the items aligned to the objectives. Questionnaires should be structured to include both Likert-type scale responses and open-ended

responses. Questions to be used include, but are not limited to the following. Additional questions may be recommended.

(a) For the objectives under each content area, did the items cover the topics identified by the objective? If not, what topics were not assessed?
(b) For the objectives under each content area, did the items cover the most important performance (DOK levels) you expected by the objectives? If not, what performance was not assessed?
(c) What is your general evaluation of the alignment between the content areas and the assessment: Highly Aligned; Moderately Aligned; Minimally Aligned; Not at all Aligned? If less than moderately aligned, what are some of the overlapping or non-overlapping features of the assessments that caused you to reach this determination?
(d) What additional comments do you have about the alignment between the assessment and the framework?

Question (a) determines if an important part of an objective/element/skill was not assessed in any way. It is possible for an assessment to have items that only partially target the full intent of an expectation. For example, a mathematics objective may expect students to represent real numbers using exponents, scientific notation, absolute values, graphs, and the number line. However, if the assessment had items that only targeted the use of the scientific notation, then the objective would only have been partially addressed. The typical coding scheme for the Webb method does not require that panelists indicate the degree to which objectives are only partially addressed, and that will be changed as a part of Task 3 to collect this information for the NAEP content alignment studies, along with the rationale of each panelist for this judgment. Coding for partial coverage will help to maximize information about the relationship between the two assessments, particularly since mapping one assessment to another assessment's framework will likely yield several partial hits. The intent of the first debriefing question is to have the panelists identify parts of objectives that were not assessed in any way, such as the use of exponents in this example.

The second question (b) is similar to the first debriefing question, but it seeks to ascertain if panelists judge that the assessment is targeted to the performance identified by the objective. Having panel members give their overall evaluation of the alignment between the assessment and the assessment framework provides a holistic judgment by people who have just finished thinking very deeply about the relationship of the two. This evaluation provides more detailed information to enhance the interpretations of the alignment data. This evaluation is not intended as an indicator of the validity of the assessment instrument.

Task 10 Facilitators Determine Discrepancies and Panelists Adjudicate: After all panelists have completed mapping all of the NAEP items to the NAEP framework, the group facilitator should review the codings from the group and conduct an adjudication process for discrepancies in either assigning items to objectives/elements/skills or DOK levels to items. Discrepancies that should be discussed are those items that have not been assigned by more than half of the panelists to the same objective/element/skill or items that have been assigned to three different DOK levels or to two non-contiguous DOK levels.

Content complexity is a continuum. The Webb alignment process uses the average DOK among the panelists for analysis. It is reasonable for panelists to assign adjacent DOK levels to an item indicating that the complexity of an item is probably between a DOK level 1 and a DOK level 2 or between a DOK level 2 and a DOK level 3. However, if some panelists are assigning a DOK level 1 (recall or recognition in mathematics) to an item while others are assigning a DOK level 3 (strategic thinking) to the same item then this difference requires discussion. After discussing these items, panel members may change their item codes, but it is not necessary to change if they feel strongly that their original judgments were correct. The adjudication process could reveal a more appropriate objective/element/skill for an item than the panelist initially selected, or it could reveal that the panelist made an error in recording his or her coding for an item.

Reports should be reviewed to determine discrepancies between the replicate panels. This review should begin only after both groups have completely finished coding the assessment items to the assessment framework and after the adjudication process. The summary reports for each of the four alignment criteria should be used to determine if the two groups are in agreement. Under Categorical Concurrence, the average number of items assigned to each content area for mathematics and aspects of reading by text type for reading should be reviewed. Under the Depth-of-Knowledge Consistency, the average percentage of items for each content area that were below the DOK level of the assigned objective/element/skill, at the DOK level, or above the DOK level should be reviewed. Under Range-of-Knowledge Correspondence, the percentage of objectives/elements/skills that had at least one corresponding item should be reviewed.

For Balance of Representation, any index value lower than .7 should be reviewed because an index value below .7 would indicate that the majority of items were coded to only one or two objectives. The balance-of-representation criterion is used to indicate the extent to which one "knowledge" expectation is given more emphasis on the assessment than another. This index only considers the "knowledge" expectations for a standard that has at least one hit—i.e., one related assessment item per expectation. The index is computed by considering the difference in the proportion of expectations and the proportion of hits assigned to each expectation. An index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a standard are equally distributed among the expectations for the given standard. Index values that approach 0 signify that a large proportion of the hits are accounted for by only one or two of the expectations. If most items relate to one expectation and only one or a few items relate to the remaining expectations, this would be described as a unimodal distribution and the index value would be less than 0.5. A bimodal distribution would have an index value of around 0.55 or 0.6. Index values of 0.7 or higher indicate a relatively even distribution of items across all of the expectations. An index value of 0.7 or higher is recommended as the target criterion for balance-of-representation. Index values between 0.6 and 0.7 indicate the balance-of-representation criterion has only been "weakly" met.

Any differences between the two groups of panelists of more than five percentage points, should be investigated further. This criterion has emerged from numerous studies as a good indicator of the level of agreement, or lack thereof, which signals the need for further evaluation and explanation. If the results are within these margins, the results for the two groups will be deemed to replicate judgments. If the differences are greater, the standard agreement tables and the item agreement tables should be examined to identify the group differences in the mapping of

items to the objectives/elements/skills. The group facilitators should first try to resolve any large discrepancies by reviewing documentation of panelists' opinions collected throughout the process. Facilitators will identify areas of disagreement to be discussed by the combined group of panel members for the subject. If the discussion does not lead to common agreement between groups, the differences will be resolved by the two group facilitators.

Task 11 Map Two Forms of the Pexam to the NAEP Framework: The same procedures should be followed in mapping items from the Pexam to the NAEP assessment framework as were followed for Task 8 (mapping the NAEP assessment to the NAEP assessment framework).

Task 12 Panel Members Respond to a De-Briefing Questionnaire: Panel members should individually respond to a small set of questions after the Pexam items have been coded to the NAEP framework (see Task 9).

Task 13 Facilitators Determine Discrepancies and Panelists Adjudicate: See Task 10 for a description of identification of discrepancies between panelists in coding Pexam items to NAEP framework and of panelists' participation in a discussion and adjudication process.

Task 14 Panelists Identify Differences between Assessments in Final Debriefing for Mapping to NAEP Framework: The mapping of the assessment items to the NAEP assessment framework will conclude with panelists individually responding to debriefing questions regarding each assessment. Questionnaires will be constructed to elicit responses to these questions, as well as to document the level of understanding, confidence, and comfort with which panelists performed the tasks. Panel members should respond to questions regarding the following aspects of the alignment of items to the NAEP framework. Additional questions may be recommended.

(a) What were major differences between the two assessments in item types, content coverage, and complexity of items relative to the NAEP framework?
(b) 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 moderately, and who perform poorly?
(c) What similarities and differences were identified between the two assessments?

Tasks 15 Assign DOK levels to Pexam Assessment Framework Expectations: Tasks 7-14 involve panelists mapping the item pools of the two assessments to the NAEP framework. Beginning with Task 15, panelists are repeating the same tasks with the Pexam framework. The tasks for this alignment can vary by major content topics included, the structure of the content, the level of specificity (grain size), and the type of performance expressed. Mapping the Pexam items and the 2009 NAEP item pool to the Pexam assessment framework will produce another view of the alignment between the two assessments. Each assessment framework is a representation of a domain of knowledge. The extent to which the mappings of two or more assessments to a common domain of knowledge are similar will help to determine the degree of alignment. Mapping the NAEP and the Pexam items directly to the framework of each assessment will reveal similarities and differences in the DOK levels of items in relationship to the framework, range, and balance as described in the section above on determining the degree of alignment using the four alignment criteria. The bi-directional alignment design maximizes

information regarding the alignment between two assessments, i.e., the degree and nature of both overlap and non-overlap between the two assessments.

The consensus process used by panelists to assign DOK levels to the NAEP assessment frameworks will be used in coding items to the Pexam assessment framework. It is likely that the Pexam assessment frameworks will not have the same level of detail as the NAEP assessment framework, so the amount of time required for this part of the process will be different. Reaching consensus on the DOK levels of objectives is still important to facilitate discussion of the framework's objectives.

Task 23 Analysis of Alignment Data: After the content alignment panels have completed their work, data should be analyzed to describe the proportion of the objectives in each the NAEP and Pexam assessment frameworks by DOK levels. The two frameworks should be compared on the proportion of objectives distributed across the different levels of complexity. This comparison should be made for each content area and general topic underlying a content area. The data on the DOK levels of the objectives in the two assessment frameworks should be interpreted in the context of the comparative content analysis of the two frameworks and test blueprints completed in Task 4. In the comparative content analysis (Task 4), an expert is to produce information on the alignment of two assessments, based on the framework documents and test specifications. The data collected from the content alignment panelists will produce information based on the actual assessment items. The analysis in Task 4 is at a higher level and reflects the "intended" assessment, whereas the analysis by alignment panelists is at a more detailed level and reflects the "actual" assessment—how the framework was operationalized by the pool of assessment items.

It is possible that items match a sub-area of an assessment framework but not the next more detailed level of organization called for by the framework, such as an "objective." In that case, a "generic objective" is identified for coding the item within a specific sub-area. If two or more panel members assigned an item to a generic objective (i.e. subtopic or content area for mathematics NAEP and text feature or literary type for reading NAEP), the items should be listed for evaluation. Generic objectives indicate the absence of complete alignment. A large number of items mapped to a generic objective indicates holes in the assessment framework—an issue of granularity for which the items only target the general ideas expressed by the framework area and not the explicit content described by the objectives. To the extent possible, such gaps should be identified in advance by the alignment contractor so that discussions can be conducted on site with panelists regarding these items and potential gaps.

Data must be recorded for each panel member to report the number of items coded to each objective/element/skill under a content area. Data analyses should include computations of averages for the different alignment criteria across panelists including the number of items assigned to an objective/element/skill and the DOK level of the item in relationship to the DOK level of the assigned objective/element/skill. Data recorded by individual panelists are averaged across panel members to produce the average number of items coded to an objective/element/skill. The variables to be computed include:

- Categorical Concurrence:
- Frequency of items by:
  Content area
  Subtopics/reading text features and literary types
  Objectives/element/skill
  Other content area
- Depth-of-Knowledge Consistency
- Frequency of items by DOK level within:
  Content area
  Subtopics/aspects/context
  Objectives/elements/skills
- Range-of-Knowledge Correspondence
- Percentage of objectives/elements/skills under a content area with one or more items
- Balance of Representation
- Balance index value for each content area

The values of each of these variables should be compared between the two assessments as mapped to the NAEP assessment framework and the two assessments as mapped to the Pexam assessment framework. The software analysis package (e.g. WAT) should be used to produce a cross-study data table that creates a table with items from each assessment mapped to the same objectives/elements/skills. The information recorded in the cross-study tables can be used to compare the distribution of items by objective for the two assessments (Exhibit 4). The example illustrated in Exhibit 4 shows that the two forms are very comparable on the Categorical Concurrence alignment criterion for Standard M.S.6.1 because the total number of items on each form assigned to each objective is very similar (only varies by one item for each objective).

Objective	Group DOK Consensus	NAEP Item ID (Freq Coded)					Pexam Item ID (Freq Coded)				
M.S.6.1	2										
M.S.6.1	2										
M.O.6.1.1	1										
M.O.6.1.2	2	8-(6)	9-(3)	40-(6)			39-(6)	41-(3)			
M.O.6.1.3	1										
M.O.6.1.4	3	13-(2)	22-(6)	39-(6)	41-(2)		15-(4)	17-(4)	32-(6)	38-(6)	
M.O.6.1.5	2	19-(6)					26-(5)	35-(6)			
M.O.6.1.6	1	6-(5)	27-(4)				2-(6)	31-(6)			
M.O.6.1.7	2	28-(6)	33-(5)	37-(6)	43-(6)		1-(3)	20-(6)	42-(4)		

Exhibit 4: Sample cross-study table produced by WAT contrasting items from NAEP and Pexam mapped to the same objective by number of panelists (given in the parentheses).

Additional data analyses should be produced that show for each item on each assessment analyzed the objective/element/skill as mapped by each panelist. These data should be used to determine the consistency among panelists in mapping an item to an objective/element/skill and

the number of objectives to which each item was mapped. These data should be used to assess the breadth in content targeted by specific items (items assigned to multiple objectives/elements/skills) and the level of interrater agreement in assigning an item to objectives/elements/skills. Given appropriate training and understanding on the part of the panelists, a lack of agreement is most likely due to overlapping objectives/elements/skills in the assessment framework and/or more robust items that provide students an opportunity to apply a number of different approaches to answering the item correctly. Panel members' notes on specific items are used to identify the source of low interrater agreement, if that is revealed.

Information from the framework content analysis should be used to describe the format of each assessment to provide a context for the analytic data generated from the analysis of the assessments by the panels. For example, the item formats used in each assessment should be described. This discussion should include the number and proportion of items included on an assessment that targeted specific content areas. If items have different score points, the comparison should be weighted to reflect this fact. For example, constructed response items in NAEP typically have a higher number of possible points than multiple choice items. Information that should be reported includes, but is not limited to, the following.

- Item format
  - Proportion of multiple-choice
  - Proportion of constructed response
  - Proportion of extended response
- Item context
  - Passage characteristics

Alignment Determination

The content alignment between the NAEP assessment and another assessment should be established by comparing the values under the different alignment criteria for each of the assessments. These data will be presented to the Governing Board for use in evaluating and reporting results of other studies in the 12th grade preparedness research program. Working with technical experts, the Board will determine quantitative criteria to be applied to the results for determining the extent to which two assessments are aligned. Values to be considered in determining the extent to which two assessments are aligned include:

(a) The number and proportion of items that map to each content area
(b) The DOK levels of the items within each content area
(c) The proportion of objectives under a content area targeted by items
(d) The relative emphasis to subtopics and objectives under a content area
(e) The structure of knowledge represented in the assessment framework for each assessment as indicated by the format of items, the item context, and other test characteristics

All of these factors need to be considered together and reported to determine the *degree* of alignment. Of course, some factors are more important than others, and the relative weights to assign to these factors will need to be determined by the Board.

Panelists

A replication of the alignment, two groups conducting the alignment concurrently, is required to strengthen the confidence in the findings. Porter, et al (2008) report that five raters provide the requisite level of reliability in a content alignment study. For the NAEP studies, however, two groups of six -eight content experts each are recommended for each subject. (The recommendation is that 8 panelists be recruited for each replicate panel to help ensure that a minimum of 6 panel members are available in the event of last-minute attrition.) As discussed above, the two groups will conduct the analysis simultaneously so that results can be compared and differences adjudicated on site. Conducting such a replication at two separate times would likely produce some variation in results, most likely at the item level rather than with the alignment criteria. The timing logistics would then make it difficult to determine the reasons for these variations. Further, previous experience indicates that panelists develop decision rules for their coding when faced with assessment frameworks that have over lapping objectives or when the objectives in an assessment framework lack clarity. These issues increase the importance of holding replicate panels simultaneously; having two groups independently, but at the same time, conduct the analysis will help reveal such decision rules and determine how these rules impact the results. To the extent possible, potential ambiguity will be identified in advance and panelists will be trained to code items in a consistent way. However, strict decision rules should be avoided as these may discourage the use of panelist expertise.

The experts for a content area who will serve as panel members should be selected because of their deep knowledge of the subject matter (mathematics or reading) and experience in analyzing curricula and assessments. Caution must be exerted in selecting persons for the panel to assure there is no bias with regard to any one of the assessments to be analyzed. Fifty to sixty percent of the panelists (in each of the two replicate panels) should come from post-secondary activities relevant to the Pexam (e.g. mathematicians, mathematics educators, language arts professors, and reading educators). Examples of individuals from the secondary education sector to serve as panelists include curriculum coordinators, content area assessment specialists, state content consultants, and high school teachers in the subject area. Intimate knowledge of each assessment should be equally represented by panelists on each panel. Individuals who have participated in other alignment studies are eligible to serve on these panels. Persons who are employed by commercial testing companies are not eligible to serve as panelists. Geographic regions and racial-ethnic groups should be represented proportionally to assure diversity of the group of panelists. Content expertise and knowledge, however, should be the primary criterion for selection of panelists.

Quality Control

Having two groups of panelists conduct the alignment analysis simultaneously for each subject will provide evidence of the reliability of the findings and increase confidence in the final results. The two groups should be trained together so the training is standardized, but the alignment analyses should be conducted independently.

As mentioned earlier, two forms of each Pexam assessment and the entire 2009 NAEP item pool for each subject should be analyzed. It is assumed that the Pexam assessment forms are parallel with very little variation in the distribution of items among content topics and by item format. Forms of the Pexam can be analyzed for comparability so that the forms selected for use in the alignment analysis are similar. The difference in the number of items between the 12^{th} grade NAEP (approximately 200 for mathematics) and a form of the Pexam (probably less than 100 items) is not critical in determining the alignment between the two assessments because results are primarily reported by proportion of items on the assessment. When the two forms are analyzed, four panelists should start analyzing one form first with the other panelists starting with the other form so that the two forms are reviewed in a different sequence.

Timeline for Conducting the Alignment Analysis

The estimated time required for conducting the alignment analysis from the approval of the proposal to submitting the final report is seven months. This time span could be compressed if some of the preparation tasks are implemented concurrently. The alignment study, data analysis, and report writing will require about three months to complete. Careful review of the report is important, and ample time must be allotted for this part of the process.

The major tasks and the estimated time for each task are listed below. Time estimates are stated as the duration needed to complete a task. Some tasks can be accomplished simultaneously.

1. Set time, location, and venue for the analysis (4 weeks)
2. Identify group facilitators and panelists (4 weeks)
3. Identify materials and equipment needed for the analysis (2 weeks)
4. Enter NAEP framework into analysis software (1 week)
5. Print and compile materials needed for coding (3 days)
6. Alignment Study (1 week)
7. Analyze data by tabulating variables for alignment criteria and comparing values between tests (2 weeks)
8. Incorporate qualitative information on tests and NAEP (2 weeks)
9. Draft report (2 weeks)
10. Review report (4 weeks)
11. Final report (4 weeks)

Logistics for the Study

The facility selected for the study must accommodate the use of a computer with appropriate network and/or internet access for each panel member. A general session room will also be required along with one additional break out room for one of the replication panels. A room set classroom style for ease of computer use is recommended for the panelists. An additional room to store materials and to conduct analyses of results, review questionnaires, and so forth is recommended.

The process and analysis will be greatly facilitated if software specifically designed for conducting alignment studies is used, such as the Web Alignment Tool (WAT). The WAT can be used by panelists to enter data and by staff for analyzing data and producing data tables. The computers should be hard-wired to a server rather than using a wireless. Most conference hotels can accommodate the requirements for this sort of study.

The Governing Board staff will acquire secure materials for NAEP and the other test programs and facilitate acquisition of other necessary assessment materials for the study. The alignment study contractor will be responsible for identifying necessary materials and for maintaining the security of all test materials.

Process Leadership

One person, the project director, should have ultimate responsibility for conducting the alignment study and exercising leadership on site. This person should be identified by the alignment contractor and serve as the primary contact person for Governing Board staff. The project director should be responsible for identifying the specific procedures to be used and should have overall responsibility for assuring that all tasks are completed on time and according to the agreed-upon study design. The project director should be responsible for overseeing the data analysis and report writing and for the production of all contract deliverables associated with the content alignment study.

The alignment contractor should clearly identify the staffing requirements for the project and who will be responsible for conducting each of the tasks. At a minimum the project director should be assisted by two subject matter group facilitators for each of mathematics and reading. One of the subject matter group facilitators should have responsibility for training all of the panelists in the content group together. There will be a facilitator for each alignment panel group when the replicate groups are formed.

The group facilitators will need to be content experts who are well versed in the alignment analysis process. The group facilitators should be experienced in training panelists in the alignment methodology and facilitating the alignment process.

Reliability, Panelist Agreement, and Replication of Results

Panelists will engage in two major judgments when coding items to an assessment framework: the assignment of a DOK level to a test item and the assignment of an item to an objective(s) under the assessment framework. For both of these judgments, consistency among panelists is critical to meaningful findings. An average measure intraclass correlation (Shrout and Fleiss, 1979) should be used as one measure of the panelists' agreement in assigning DOK levels to items. The average DOK level assigned to an item by the panelists should be used as one of the variables in the analysis. The pairwise comparison is a more stringent statistic and should also be used. If variance in assigning DOK levels to items among panelists is low, then computation of the intraclass correlation is inappropriate. The pairwise comparison provides a more meaningful

measure of agreement in this case. Pairwise comparisons can also be used as one measure to judge the agreement among panelists in assigning items to objectives, elements, subtopics, and the content area.

Alignment judgments by panelists will not be in complete agreement, and several sources of variance should be analyzed. The effectiveness of the training, the composition of the alignment panel, the depth of discussion among the panelists, and the sequential order of aligning different forms of assessments can all generate variations in the final results and impact the apparent degree of alignment between two tests. Conducting a replication study will add confidence to the findings, identify information that is consistent across the groups, and identify inconsistencies across groups and studies. The most rigorous replication study design would require a second panel to independently conduct the analysis in its entirety. The recommended design using concurrent replicate panels allows adjudication of differences through discussion between members of the two panels. If the variation between the two panels appears to be random and minor, then the results from the two panels can be averaged or aggregated.

Materials

Materials from different sources are required for the study. Copies for each panelist, for group facilitators, and for observers will be needed.

Required NAEP Materials

Assessment frameworks and test specifications for reading and mathematics
- NAEP 12th grade item pool for each subject
- Scoring guides and scoring rubrics for each item
- Anchor papers to represent each score point for each constructed response item
- Point value assigned to each item

Required Pexam Materials
- Test blueprints
- Test objectives
- Two test forms of each test to be analyzed
- Scoring guide and scoring rubrics for each item
- Anchor papers to represent each score point for each constructed-response item
- Point value assigned to each item

Additional Materials
- Training materials (DOK definitions, illustrative items)
- Software for data entry and analysis
- Presentation materials
- Evaluation forms
- Computers, printers, photocopiers

Evaluations by Panelists

Panelists will be given evaluation questions after the training, after major tasks, and at the end of the study. The questionnaires will include both structured response items (Likert-type scales) and open-ended questions. These questions will focus on panelists' evaluation of the following aspects of the study:

1. Training and instructions
2. Materials, both advanced and on site
3. The alignment process, including the qualifications of panelists, composition of panels, alignment criteria, coding of items, quality and quantity of information provided, and adjudication procedures
4. Procedures for data communications, especially the ease of using the software
5. Logistics, including meeting facilities, agenda, travel arrangements

Evaluation questions after the training include:

1. How well do you feel the training prepared you to apply the Depth of Knowledge Levels?
2. How well do you feel the training prepared you for the adjudication process?
3. Overall, how well did the training prepare you for the alignment process?

Evaluation questions to be answered at the end of the study:

1. How well did the process capture the content similarities of the assessments?
2. How well did the process capture the content differences between the assessments?
3. To what degree was the pair of assessments aligned?
4. Considering the items in each assessment, how did the assessments differ and how were they the same?

Reports

An interim report is to be presented regarding the expert comparison of the two assessment frameworks conducted as Task 4. This report should present the results of the evaluation and specify how the information will be used to structure the data entry and reporting software, to identify and eliminate ambiguous objectives or other framework aspects, to train panelists for coding items, and to report results of the alignment study. Any other aspects of the alignment study that should be addressed must be identified in this report. An overall evaluation of the alignment of the two assessments, based on the comparative analysis of frameworks, should be stated.

A comprehensive report is to be prepared to describe the methodology used and the results of the alignment between each NAEP assessment and the Pexam assessment. The methodology section of the report is to describe the qualifications of the group facilitators and panelists, the structure of the assessment framework used in the analyses, the training of panelists, the alignment criteria used in the analysis, and the coding procedures. The comparative analysis of the assessment frameworks for the assessments should provide the context for reporting the findings on the alignment between the assessments. The results section will summarize the DOK levels assigned

to the objectives of the assessment framework, the description of the results of mapping each test to the assessment framework, the degree of alignment between each NAEP subject item pool and each Pexam. The conclusions will delineate the parts of the two assessments that are aligned and the parts that are not aligned. Within each category, variability in level of alignment and non-alignment should be identified.. An Executive Summary will present the major points and conclusions from the report. Data tables produced from the analysis are to be included as appendices to the report.

References

Ananda, S. (2003). Rethinking issues of alignment under No Child Left Behind. San Francisco: WestEd.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), p21-29.

Council of Chief State School Officers (2006). Aligning Assessment to Guide the Learning of All Students: Six Reports on the Development, Refinement, and Dissemination of the Web Alignment Tool. Washington, D.C.: 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.

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Hansche, L. (2000). *State Standards and State Assessment Systems: A Guide to Alignment.* Washington, DC: Council of Chief State School Officers.

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 (2008a). Mathematics Framework for the 2009 National Assessment of Educational Progress. Washington DC: Author.

National Assessment Governing Board (2008b). Reading Framework for the 2009 National Assessment of Educational Progress. Washington DC: Author.

Porter,A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31 (7); 3-14.

Porter, A. (2006). Measuring alignment. *National Council on Measurement in Education Newsletter*, 14 (4), December, 2006.

Porter, A. C., & Smithson, J. L. (2002, April). Alignment of assessments, standards and instruction using curriculum indicator data. Paper presented at the Annual Meeting of American Education Research Association, New Orleans, LA.

Porter, A.C., Polikoff, M.S., Zeidner, T., & Smithson, J.L. (2008). The quality of content analyses of state student achievement tests and content standards. *Educational Measurement: Issues and Practice*, 27 (4), 2-14.

Shrout, P.E. & Fleiss, J.L. (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 2, 420-428.

Smithson, J. L. & Porter, A.C. (2004). From policy to practice: the evolution of one approach to describing and using curriculum data. In M. Wilson (Ed.), *Towards Coherence Between Classroom Assessment and Accountability. The 2004 yearbook of the National Society for the Study of Education*. Chicago: University of Chicago Press.

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

Webb, N. L. (1997). Criteria of alignment of frameworks, standards, and student assessments for mathematics and science education. Council of Chief State School Officials and National Study for Science Education Monograph, No. 6. Madison: Wisconsin Center for Education Research.

Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states (Research Monograph No. 18). Madison: University of Wisconsin–Madison, National Study for Science Education.

Webb, N. L. (2202. Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states. Washington, D.C.: Council of Chief State School Officers.

Appendix A
Example of Agenda

Day 1: Training in General Session

1. Introductions and administrative details (one hour)
2. Training (5 hours)
   a. Purpose and importance of the study
   b. Overview of the process
   c. Training in specific tasks
3. Instructions for log in to the WAT or other software for this purpose (one hour)
4. Evaluation of training (.25 hours)

Day 2: (Parallel Replicate Panel Groups

1. Code DOK levels of NAEP framework (2-3 hours)
2. Map NAEP items to NAEP framework (4-5 hours)
3. Break (Facilitators check item coding and identify discrepancies for discussion and adjudication process)
4. Adjudicate coding of NAEP items to NAEP framework (1 hour)
5. Evaluation of NAEP items-to-NAEP framework coding (.25 hour)
6. Evaluation of process and understanding of procedures (.25 hour)

Day 3:

1. Map Pexam Form 1 to NAEP framework (2.5 hours)
2. Map Pexam Form 2 to NAEP framework (2.5 hours)
3. Break (Facilitators check item coding and identify discrepancies for discussion and adjudication process)
4. Adjudicate coding of Pexam items to NAEP framework (1 hour)
5. Evaluation of Pexam items-to-NAEP framework coding (.25 hour)
6. Evaluation of process and understanding of procedures (.25 hour)

Day 4:

1. Code DOK levels of Pexam framework (2-3 hours)
2. Map Pexam Form 1 to Pexam framework (2.5 hours)
3. Map Pexam Form 2 to Pexam framework (2.5 hours)
4. Break (Facilitators check item coding and identify discrepancies for discussion and adjudication process)
5. Adjudicate coding of Pexam items to Pexam framework (1 hour)
6. Evaluation of Pexam items-to-Pexam framework coding (.25 hour)
7. Evaluation of process and understanding of procedures (.25 hour)

Day 5:

1. Map NAEP assessment items to Pexam framework (4-5 hours)
2. Break (Facilitators check item coding and identify discrepancies for discussion and adjudication process)

3. Adjudicate coding of NAEP items to Pexam framework (1 hour)
4. Evaluation of NAEP items-to-Pexam framework coding (.25 hour)
5. Evaluation of process and understanding of procedures (.25 hour)
6. Overall debriefing (1 hour0
7. 1.5 hours debriefing across assessments
8. Evaluation of overall alignment process, evidence generated, criteria applied, and holistic conclusion regarding alignment of the assessments; recommendations regarding alignment and appropriate uses of evidence; evaluation of process and understanding of procedures (.5 hour)

Appendix B

Depth-of-Knowledge Definitions for Mathematics[2]

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 a 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 the objects or phenomenon 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 requiring mathematics information from the graph, also is 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 solely to number skills, but can involve visualization skills and probability skills. Other Level 2 activities include noticing and 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

[2] Mr. Webb has used the DOK definitions for conducting alignment studies, and he judged the definitions to be applicable to an alignment study with NAEP assessments. The definitions shall be used as described; however, recommendations may be made for approval of changes to the definitions in cases where this is necessitated by virtue of the content of the Pexam.

developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

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 Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be at 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 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 developing and proving conjectures; designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Appendix C

Depth-of-Knowledge Definitions for Reading[3]

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

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

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

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

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

[3] As for mathematics, the DOK levels for reading describe levels of content complexity that can be used for analyzing the NAEP and other reading assessments. The reading levels are based on Valencia and Wixson (2000, pp. 909-935). The definitions shall be used as described; however, recommendations may be made for approval of changes to the definitions in cases where this is necessitated by virtue of the content of the Pexam.

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

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

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

Appendix D

Example Comparing a NAEP Mathematics Subdomain with a Standard from Another Test

NAEP Mathematics	Test A Mathematics	Similarities and Differences
Number Properties and Operations		
1) Number Sense	N. NUMBER SENSE	Test A: Number sense is one of five standards. NAEP: it is subtopic under one of five content areas.
	N.1. Analyze the structural characteristics of the real number system and its various subsystems. Analyze the concept of value, magnitude, and relative magnitude of real numbers.	
d) Represent, interpret, or compare expressions for real numbers, including expressions using exponents and logarithms.	N.1.1. Students are able to represent numbers in a variety of forms and identify the subsets of rational numbers. • Exponents • Scientific notation • Absolute value • Radicals (perfect squares) • Graph on a number line	Both frameworks state students are to represent real numbers using exponents. NAEP only includes logarithms. Test A explicitly states radicals and number line graphs.
f) Represent or interpret expressions involving very large or very small numbers in scientific notation.		NAEP incorporated in Test A N.1.1.
g) Represent, interpret, or compare expressions or problem situations involving absolute values.	N. 2. Apply number operations with real numbers and other number systems.	Test A explicitly states absolute value as a form to be represented. NAEP expects specific applications of absolute value in problem situations
	N.2.1. Students are able to read, write, and compute within any subset of rational numbers. • Solve problems involving discount, markup, commission, profit, and simple interest.	NAEP more explicitly uses computation with real numbers under subtopic number operations.
i) Order or compare real numbers, including very large or small real numbers.		

	N.3. Develop conjectures, predictions, or estimations to solve problems and verify or justify the results.	NAEP as a subtopic for estimation.
	N.3.1. Students are able to use various strategies to solve multi-step problems involving rational numbers. • Explain strategies and justify answers. • Formulate rules to solve practical problems involving rational numbers. • Use estimation strategies to make predictions and test the reasonableness of the answer.	Test A includes using estimation strategies to make predictions. NAEP attends to level of accuracy and verifying results. Test A expects students to solve problems involving rational numbers. NAEP under subtopic Number Operations expects students to use real numbers.

Appendix E

The following chart includes some steps that have been useful in previous alignment studies to facilitate the consensus process.

Facilitating the Consensus Process	
1.	Read each objective aloud before discussing it.
2.	As you go through the objectives, actively solicit comments from all panel members. Pay special attention to making sure that each panel member feels involved. (Not every panelist needs to address every objective, but make sure that everyone is included in the process.)
3.	Use the print-out to call on people who coded DOK levels differently from the coding of other members of the group, and ask them to explain why they coded the objective to the particular DOK level. Be sure they use the DOK definitions to justify their answers.
4.	Once two panel members have described how they have coded an objective differently, ask a third panel member to highlight the differences between these two interpretations.
5.	Restate and summarize points of agreement and disagreement among panelists to determine if your interpretation is accurate.
6.	If there is a difference in interpretation of the objective's *terminology* or *expectations*, discuss alternatives by asking for volunteers with direct experience in applying an objective.
7.	Provide an opportunity for panelists to change their codes after the discussion.
8.	If panelists remain divided on the DOK level of an objective, focus attention on the most likely skills or content knowledge required in the objective, not the more extreme possibilities the objective might allow.
9.	The facilitator should not dominate the consensus process. Even if the facilitator has strong feelings about the DOK level of an objective, it is important to have panel members raise the points and reach agreement on level.