Considerations For Developing Test Specifications

For Common Core Assessments

Adopting Curriculum Standards—Only the First Step

A white paper from Pearson

Michael Bay-Borelli
Christine Rozunick
Walter D. Way
Eric Weisman

December 2010
Pearson, the global leader in education services, education technology and school solutions, provides innovative print and digital education materials for preK through college, student information systems and learning management systems, teacher professional development, career certification programs, and testing and assessment products that set the standard for the industry. Pearson’s other primary businesses include the Financial Times Group and the Penguin Group.

Pearson’s research reports allow the dissemination of its research and assessment-related articles before publication. Pearson reports in .pdf format may be downloaded from: http://www.pearsonassessments.com/research.
Abstract

The purpose of this paper is to describe the role that test specifications play in supporting the development of valid and reliable large-scale summative academic achievement assessments. The paper describes the relationship between curriculum standards, which are primarily guidelines for instruction, and test specifications, which are guidelines for the development of sound and effective items and tests. The paper also identifies and discusses critical content issues associated with the establishment of test specifications and offers recommendations regarding how those issues might best be addressed by the consortia seeking to convert the Common Core State Standards into 21st century assessments.

An executive summary is provided first. This summary includes details about the questions discussed for each issue and a summary of recommendations. Each issue includes a hyperlink to the full text of the discussion in the paper that follows.
Considerations For Developing Test Specifications For Common Core Assessments

Executive Summary with Recommendations

The Common Core State Standards (CCSS) were adopted to provide guidance to stakeholders regarding what students should understand and be able to do at various critical points in their education. The standards will also provide the basis for the next generation of large-scale state assessment programs. Curriculum standards alone, however, are not sufficient to define a sound and reliable large-scale assessment program; well-defined test specifications are also necessary.

State policy makers must have a reasonable degree of assurance that the assessments developed measure student learning relative to the curriculum standards. While the CCSS describe what should be taught, the test specifications that must be developed will ultimately describe how student achievement relative to those curriculum standards will be measured.

We believe that it is essential that the policy makers collaborate to address the content issues detailed in this article and to reach clear consensus regarding solutions. Successful collaboration within a state is often challenging; successful collaboration across multiple states will undoubtedly be even more challenging. Without that clear vision of the test and its parameters, it is not possible to develop stimuli and items or build tests that will fairly and reliably measure student learning.

Anticipating all of the content issues that will need to be addressed is not possible, but some of the more salient ones can be anticipated and are discussed in this article. These issues should be resolved collaboratively before developing operational test questions. The content issues address both item development and test design. A list of critical questions associated with each issue is provided, followed by discussion and recommendations.

A brief summary of the questions addressed and the recommendations offered in the article follows. For the more complete discussion, please refer to the full article.

Issue #1. Curriculum Standards

Questions
- For a given subject and grade, will all of the Common Core standards or only a subset be deemed eligible for assessment?
- If only a subset is eligible, what criteria will be used to determine ineligibility, who will make the determination, and what process will be used to make the determination?

Recommendations:
1.1 Clearly articulate and communicate the criteria used to deem certain standards ineligible for testing to all stakeholders. When making eligibility decisions, the criteria should be applied consistently across grades within a subject area and, ideally, across subject areas.

1 The term “item” should be taken throughout this article to mean the full array of assessment content used on large scale assessments, ranging from multiple-choice questions to constructed response including performance tasks.
1.2 Involve policy makers, content specialists, and educators in the decision making process. The consortia must have a clear understanding of who will be involved in formulating the decision and who will have the authority to make the final decision.

1.3 While certain skills may appear to be impossible to measure on a test, a more inclusive eligibility policy is advisable in the hopes that creativity and technology will help to provide new ways to measure student achievement for more challenging standards.

Issue #2. Test Focus Relative to Curriculum Standards

Questions
- Will all assessable standards be tested each year, or will the standards be sampled for any given test?
- If the standards are sampled, how will the sampling be determined, who will determine the sampling method, and what process will be used to determine the sampling method?

Recommendations:
2.1 Give all policy makers and knowledgeable stakeholders input into the decisions regarding which standards are tested in each administration and which standards might be sampled across administrations in a rotated fashion. These decisions have profound implications for the test blueprint, item development plans, test construction, scoring and reporting, standard setting, etc.

2.2 Involve policy makers, content specialists, and especially classroom educators in deciding whether to sample the skills assessed because of desired test length, curriculum emphasis, or the volume of standards. These “on the ground” considerations related to test specifications are critically important in constructing sound test specifications.

2.3 If a consortium chooses to sample the CCSS skills on any assessment, all policy makers should clearly articulate the rationale for that decision to stakeholders. The risk of an unintended consequence deriving from a decision to sample standards is significant—namely, that stakeholders may come to believe that something less than the entire set of curriculum standards needs to be taught. This is a risk because of the intimate relationship between assessment and instruction (the “tail wagging the dog” syndrome).

Issue #3. Roll-up/Reporting Categories

Questions
- How will the assessable standards be organized into groups of related standards for reporting student performance?
- How many standards at minimum and at maximum will be in any roll-up category?
- What labels or descriptors will be used to describe these reporting categories?
Recommendations:
3.1 The set of standards drawn into a reporting category must have a logical basis and construct cohesiveness that both content specialists and stakeholders will recognize and support. When feasible, the schema used should be consistent across grades or levels within a given subject area.

3.2 The reporting logic should speak to the requirements of a 21st century curriculum model and address the emerging need for large-scale assessments to inform judgments about the college and career readiness of students.

3.3 The number of items or tasks drawn into a reporting category must be sufficiently large to support reliable subscores. In addition, the standards that are aggregated in arriving at subscores must be reasonably stable from one administration to the next.

3.4 All knowledgeable stakeholders should have the opportunity to react to the overarching logic of the reporting categories established and to the details regarding which curriculum standards are subsumed under which objectives.

Issue # 4. Curriculum Standards Alignment—Coding

Questions
- Will each item or task developed be aligned/coded to a unique Common Core standard, or will cases exist in which items or tasks are coded into multiple standards?
- Will items or tasks also be aligned/coded to other sets of standards such as those of individual states within a consortium?

Recommendations:
4.1 A consortium should make decisions about whether items and tasks will be coded to one curriculum standard or to multiple standards (within a given set of standards or across standard sets) before item development begins.

4.2 If items are to be aligned to multiple sets of standards, a consortium must make a clear decision regarding which set of standards the items will be “written to,” because the items cannot be developed with multiple masters in mind.

Issue # 5. Content Validity

Questions
- Who will review and confirm the proposed alignments of items to the Common Core standards, and what process will be used for the review and confirmation?
- How will the state or consortium document that the cognitive complexity level of the state’s curriculum standards and the state tests are aligned?

Recommendations:
5.1 The consortia must establish and follow well-defined processes to confirm the content match between the curriculum standards and the items and to verify that the average cognitive complexity of the items matches that of the
curriculum standards. These processes should include all knowledgeable stakeholders, especially classroom teachers.

5.2 Stakeholders should be careful not to define item development guidelines that are too strict. If development guidelines are too prescriptive, they may lead to “cookie-cutter” items, which could limit the success of the assessment program in reaching its stated goals.

5.3 The items developed should be reviewed and scrutinized by knowledgeable stakeholders, including policy makers, content specialists, and classroom educators.

**Issue # 6. Item Types**

**Question**
- Which assessment item types will be used to measure student achievement?

**Recommendation:**
6.1 All policy makers and knowledgeable stakeholders must be involved in decisions about the mix of item types used in an assessment because of how significantly the mix of item types affects the overall test design, content validity of the test, item development, and the scoring and reporting of student achievement.

**Issue # 7. Stimulus Types**

**Questions**
- What types of stimuli will be used in the assessments?
- Will the passage stimuli be permissioned or commissioned?

**Recommendations:**
7.1 All policy makers and knowledgeable stakeholders, especially classroom teachers, must be involved in decisions about the types of literary selections to be used in the Reading assessments.

7.2 If stimuli with copyright protections are to be used, then legally defensible permissions procedures should be established and followed scrupulously.

**Issue # 8. Test Blueprints for Each Subject and Grade**

**Questions**
- For passage-based tests, what will be the number, type, and complexity of stimuli on the test?
- Which item types and how many items of each type will be on the test?
- How many points will each item/item type be worth?
- How many items/points will be in each reporting category? Will dual-coded item scores contribute to only one reporting category or both?
- If items are also coded to external standards (e.g., National Council of Teachers of Mathematics standards) will there be test construction requirements around those secondary standard sets?
Recommendations:

8.1 Policy makers, content experts, and the psychometrics team need to develop the test blueprint together because the test blueprint reflects content decisions and psychometric considerations.

8.2 The test blueprint should be officially adopted by the assessment program’s policy board because the blueprint is a policy document.

8.3 Although test blueprints for consortia assessments could reflect quantitative requirements relative to multiple sets of standards, this is not recommended because of the significant test construction challenges this would present.

Issue # 9. Psychometric Specifications

Questions
- What role do psychometric specifications play in the overall test specifications?
- How do psychometric specifications relate to item development?

Recommendations:

9.1 Psychometric specifications should be given careful consideration by policy makers and content developers at the earliest stages of the assessment program.

9.2 Consideration should be given to piloting items through strategies such as cognitive labs early in the item development phase. Doing so can provide information to inform item writing efforts so that the statistical characteristics of the resulting items are consistent with the expectations of policy makers and psychometricians.

Issue # 10. Test Administration Modality

Questions
- Which testing modalities (paper/online) will be used to measure student achievement?
- If online, will innovative assessment strategies be used?
- If paper, how will the use of color or other text enhancement tools (e.g., boldface or italicized text) be addressed?
- If paper and online, how will comparability be addressed?
- How will the use of student tools (e.g., dictionary, compass) be addressed?

Recommendations:

10.1 If tests are to be offered in both online and paper modes, policy makers and stakeholders will need to address practical test administration issues (e.g., the use of tools) before item development begins.

10.2 If tests are to be offered in online mode, policy makers must decide if innovative online item types will be included (e.g., computer simulations, “drag-
and-drop” items, the use of video animations as stimuli, dynamic graphing options in mathematics, etc.).

10.3 The resolution of “style” issues for paper and/or online administrations should be handled through the development of a “style guide.” The style guide should be adopted by the assessment program’s policy board. All stakeholders who review the items should adhere to the style rules.

10.4 If tests are to be offered in both online and paper modes, policy makers will need to address both the psychometric comparability issues and comparability perception issues.

**Issue # 11. Performance Scoring Models**

**Questions**
- How will constructed-response (short or long) items be scored?
- Will a holistic or analytic scoring model be used to score student writing samples?
- Will student writing samples be scored by human scorers or by computer scoring engines?
- What scoring agreement model will be used?
- How will rangefinding be conducted?
- Who will monitor the performance scoring program?
- What issues must be addressed for multiple-choice, gridded-response, and hot-spot items that are to be scored by machine?

**Recommendations:**

11.1 The scoring model used for constructed-response items and the specific details surrounding the scoring rubric must be resolved before the item specifications and the items are developed.

11.2 Policy makers should determine the number of score categories for each performance task, the specific criteria for scoring each task, the number of scorers who will evaluate each response, the agreement model to be used by readers to determine student scores, and the monitoring of scorers that will be used to achieve validity and reliability before the constructed-response items are developed.

11.3 If machine-scored items such as gridded items, hot spots, and drag-and-drops are used, additional scoring considerations must be addressed before item development begins.

**Issue # 12. Accessibility Guidelines**

**Questions**
- How will universal design (UD) principles be best incorporated in item and test development requirements or item evaluation criteria?
- How will items be reviewed or evaluated relative to accessibility standards?
Recommendations:
12.1  Policy makers and content specialists should work collaboratively with all knowledgeable stakeholders, especially classroom educators and special-needs experts, in prioritizing the order in which UD principles are applied during the design and development of both the test content and testing interfaces used in the assessment program.

12.2  The UD principles agreed upon should be adopted by the assessment program’s policy board. The principles should be shared with the stakeholders who will review items, and UD training should be provided for all participants who will review and approve items.

12.3  State policy makers should establish guidelines for the accessibility of items and tests for special populations before item development begins.

Issue # 13. Item and Stimulus Review

Questions
- What criteria will be used to review items and stimuli?
- How will review meetings be conducted?
- Who will participate in such meetings, and how will participants be selected?
- At which meetings will edits be permitted, and how will the final set of edits to be implemented be established?

Recommendations:
13.1  The item and stimulus review criteria for the assessment program should be established early in the item and stimulus development process, and the final version of the criteria should be shared with all individuals and groups who will be conducting such reviews.

13.2  The review criteria should be developed with input from all stakeholders and may be adopted by the assessment program’s policy board.
The Role of Test Specifications in the Development of High-Quality Academic Achievement Assessments

Introduction
State education departments and state legislatures in all 50 states have for many years adopted sets of standards for use in instruction and assessment. More recently, the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA) led an effort to develop a set of evidence-based standards in English Language Arts and Mathematics essential for college and career readiness in a 21st century, globally competitive society. These Common Core State Standards (CCSS) were adopted in an effort to provide clear guidance to stakeholders (students, parents, teachers, school administrators, and policy makers) regarding what students are expected to learn, what they should understand, and what they should be able to do at various critical points in their educational journey. These curriculum standards will not only serve to guide classroom instructional practice in participating states, but will undoubtedly provide the basis for the next generation of large-scale state assessment programs. Curriculum standards alone, however, are not sufficient to define a sound and reliable large-scale assessment program; well-defined test specifications are also necessary.

Since the advent of the No Child Left Behind Act (NCLB) and the subsequent implementation of rigorous state accountability programs state assessment programs have increasingly taken on “high-stakes” characteristics for some students and have especially done so for teachers, school buildings, and districts. These changes mean student results on large-scale assessments that include the CCSS must be supported well by evidence of validity and reliability. State policy makers must have a reasonable degree of assurance that the assessments developed will measure student learning relative to the curriculum standards. While the CCSS describe what should be taught, the test specifications that must be developed will ultimately describe how student achievement relative to those curriculum standards will be measured.

Given this onus, the test specifications must be developed deliberately and thoughtfully with input from numerous stakeholders before developing any assessment items or tests. The test specifications will delineate the underlying architecture of the new assessments and characterize the content boundaries of the assessment program. A clearly articulated set of test specifications contribute to improved reliability and validity of the assessment instruments, thereby setting the groundwork for reliable standard setting.

Whether an assessment program is being designed for one state or for a consortium of states, it is essential that the policy makers collaborate to address the content issues detailed in this article and to reach clear consensus regarding solutions. Successful collaboration within a state is often challenging; successful collaboration across multiple states will undoubtedly be even more challenging. Without that clear vision of the test and its parameters, it is not possible to develop stimuli and items or build tests that will fairly and reliably measure student learning.

Test Specifications
The development of a large-scale multi-state assessment program involves thoughtful deliberation by all stakeholders about numerous issues associated with the design of the program. Anticipating all of the content issues that will need to be addressed is not possible, but some of the more

2 The term “item” should be taken throughout this article to mean the full array of assessment content used on large scale assessments, ranging from multiple-choice questions to constructed response including performance tasks.
salient ones can be anticipated and are discussed in this article. These issues should be resolved collaboratively before developing operational test questions. This article assumes that the CCSS consortia have determined which subjects and grades will be tested and has adopted curriculum standards for each discipline to be assessed.

These issues should be resolved collaboratively before developing operational test questions.

This article focuses on the content of the tests—the match between what is mandated to be taught and how student achievement relative to that mandated instruction will be measured. The content issues address both item development and test design. A list of critical questions associated with each issue is provided, followed by discussion and recommendations.

**Test Design and Item Development Issues**

The design of the tests used in an assessment program is influenced by numerous factors. Test designers must consider factors such as the relationship between the assessable curriculum standards and the test coverage, the modalities in which the test will be administered, the organization of the test into reporting categories, the possible need to align the tests with multiple sets of curriculum standards, and many other factors. Preeminent among these factors, however, is the purpose of the assessment program.

Policy makers must have a clearly defined vision regarding what the test is designed to measure and the purposes for which the resulting data regarding student performance will be used before test design and item development issues can be meaningfully addressed. For example, an assessment program developed to provide classroom teachers information about an individual student’s achievement for the purpose of informing instruction would have a very different structure than would an assessment program designed to inform judgments about whether students have demonstrated sufficient mastery of skills to meet the passing standards for a given course.

Policy makers must have a clearly defined vision regarding what the test is designed to measure and the purposes for which the resulting data regarding student performance will be used before test design and item development issues can be meaningfully addressed.

The purpose of the test clearly influences test design, item development, and how student performance will be reported. For example, if the purpose of the assessment is to inform classroom instruction, results are usually reported at a far more granular level than the results of a summative assessment. A formative assessment is typically designed to focus in greater depth on a narrower band of content than a high-stakes end-of-course assessment. The interplay between test purpose and required reporting formats has a profound impact on the content and structure of the test, and these issues should be considered and resolved before any other issues are addressed.
The thirteen issues discussed in this article are all critical to the design of the assessment program. They are not listed in priority order or in the order in which they must be addressed because there is substantial interaction among all of them. For example, no discussion about a test blueprint could proceed without consideration also being given to the assessable standards, administration modality, item types, reporting categories, and scoring models. Policy makers and knowledgeable stakeholders may find it desirable to address the following issues in isolation, but because of the interplay between these issues they are encouraged to maintain close communication among discussion groups.

1. Curriculum Standards  
**Questions**
- For a given subject and grade, will all of the Common Core standards or only a subset be deemed eligible for assessment?
- If only a subset is eligible, what criteria will be used to determine ineligibility, who will make the determination, and what process will be used to make the determination?

**Discussion** – While state curriculum standards typically mandate that all of the skills in the standards should be taught in classrooms, large-scale assessments cannot always feasibly measure student achievement relative to every curriculum standard. For example, the Common Core ELA standards ask students to “conduct sustained research projects” and to “write routinely over extended time frames.” Although “through-course” assessments, such as those envisioned by the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium could possibly address such standards, these standards would be virtually impossible to address in an end-of-year summative assessment. Similarly, the CCSS in high school Algebra ask students to show that they “understand that rational expressions form a system analogous to the rational numbers, closed under addition, subtraction, multiplication, and division by a nonzero rational expression.” An intellectual abstraction at this cognitive complexity level would be difficult to assess in any standardized assessment.

The Common Core standards frequently include skills that address higher-order thinking (e.g., the ability to analyze complex situations and formulate solutions to problems). The ability to measure student performance relative to such skills is challenging and is predicated to some extent on the type of assessment contemplated. An assessment limited to a paper administration using only multiple-choice items might be less likely to address some of these skills than would a computer-administered test using innovative technologies.

Determining the skills that will be deemed eligible for assessment in the new consortium-based assessment programs will be a critical first step. As noted, this determination is driven in part by whether the tests under consideration are formative or summative and by the testing formats selected. And the formats to be used can be influenced by any fiscal constraints placed on the assessment program because the increased costs associated with both the development and scoring of more authentic forms of assessment items are significant. On the other hand, the information obtained relative to student mastery of such skills can be far greater with richer items and testing formats.
Recommendations:

1.1 Clearly articulate and communicate the criteria used to deem certain standards ineligible for testing to all stakeholders. When making eligibility decisions, the criteria should be applied consistently across grades within a subject area and, ideally, across subject areas (to the extent that doing so makes sense). The process used to determine the eligibility of curriculum standards for assessment is critical if all stakeholders are to feel ownership of the testing program.

1.2 While it may not be feasible to garner input from all stakeholders, efforts should be made to involve policy makers, content specialists, and educators in the decision making process. The consortia must have a clear understanding of who will be involved in formulating the decision and who will have the authority to make the final decision.

1.3 While certain skills may appear to be all but impossible to measure on a test, a more inclusive eligibility policy is advisable, in the hopes that creativity and technology will help to provide new ways to measure student achievement for more challenging standards.

2. Test Focus Relative to Curriculum Standards

Questions

- Will all assessable standards be tested each year, or will the standards be sampled for any given test?
- If the standards are sampled, how will the sampling be determined, who will determine the sampling method, and what process will be used to determine the sampling method?

Discussion – After determining which CCSS are eligible for assessment, a consortium must also determine if it is feasible and/or desirable to test student performance on all of the eligible standards on each test. A number of factors may influence this decision.

One factor may be the sheer volume of the standards for a given subject relative to the desired test length. For example, the CCSS for high school Mathematics contain more than 175 individual skills statements (divided across six conceptual categories, not by specific course titles). If there was a desire to have one high school Mathematics test, and it contained at least one item per skill, that would result in a test of at least 175 items, and if field-test items were also embedded, even longer. Such a test would be deemed by many stakeholders to be too long. Other factors may also need to be considered, such as administrative decisions regarding the number and length of testing sessions.

Another factor influencing the decision whether to test student performance on all of the eligible standards each year may be a desire on the part of a consortium to emphasize certain eligible skills on the assessment (e.g., by testing them every year, and possibly with multiple items) and de-emphasize other skills (e.g., by testing them in alternate years). By identifying a subset of skills within the mandated curriculum as “core” or “essential,” the new tests might reflect a growing national consensus to focus assessments on skills essential for success in future grades or courses.
The state or consortium may determine that assessing student performance on a sample of the skills within some reporting categories is sufficient. For example, Reading reporting categories might include Vocabulary, Comprehension, Literary Analysis, and Critical Thinking. The group might decide to test five of the eleven skills in the Vocabulary category each year, but not the same five skills each year because the group believes a valid and reliable Vocabulary subscore in Reading can be determined by sampling these skills. Or the consortium might decide to test all Comprehension skills every year to emphasize those skills, but to sample Vocabulary skills each year.

**Recommendations:**

2.1 Give all policy makers and knowledge stakeholders input into the decisions regarding which standards are tested in each administration and which standards might be sampled across administrations in a rotated fashion. These decisions have profound implications for the test blueprint, item development plans, test construction, scoring and reporting, standard setting, etc.

2.2 Involve policy makers, content specialists, and especially classroom educators in deciding whether to sample the skills assessed. These “on the ground” considerations related to test specifications are critically important in constructing sound test specifications.

2.3 If a consortium chooses to sample the CCSS skills on any one assessment, all policy makers should clearly articulate the rationale for that decision to stakeholders. The risk of an unintended consequence deriving from a decision to sample standards is significant—namely, that stakeholders may come to believe that something less than the entire set of curriculum standards needs to be taught. This is a risk because of the intimate relationship between assessment and instruction (the “tail wagging the dog” syndrome).

3. **Roll-up/Reporting Categories**

**Questions**

- How will the assessable standards be organized into groups of related standards for reporting student performance?
- How many standards at minimum and at maximum will be in any roll-up category?
- What labels or descriptors will be used to describe these reporting categories?

**Discussion** – Categorizing the assessable curriculum standards for any subject and grade into related standards is a critical step in designing the assessment. Categorization allows judgments to be made about student performance at skill aggregation levels of sufficient size to be statistically stable. The reporting categories established will help shape the test blueprint and will determine how student performance is reported to stakeholders.

Typically, in a summative large-scale assessment there will be at most one or possibly two items assessing a particular discrete skill in the standards. Any conclusions made about student achievement relative to that skill based on performance on such a small number of items would not be reliable. But if
performance on a sufficiently large group of related items is aggregated, a score or related information (e.g., a performance category classification) may be provided. Such an aggregation of skills is usually referred to as a testing subscore or a reporting category. The set of skills drawn together for reporting purposes is usually given a descriptive label that quickly communicates to stakeholders the set of related skills being addressed.

For example for 4th grade, the Mathematics reporting categories might be built around the CCSS 4th grade domains: Numbers and Operations; Algebraic Reasoning; Measurement and Data; and Geometry; and might also include a process standard such as Problem Solving or Mathematical Reasoning. Student performance would not be reported on a single item assessing “Add and subtract mixed numbers with like denominators,” but might be reported for the larger category—Numbers and Operations (including performance on a set of related items assessing equivalent fractions, ordering fractions, and the addition, subtraction, multiplication, and division of fractions). Determining these aggregation sets, these reporting categories, is critically significant.

The organizational structure of Common Core standards may well suggest a partitioning logic. However, due to the many considerations involved, a CCSS standards domain may need to be partitioned into two reporting categories, or conversely, two standards domains may need to be combined into one reporting category.

Recommendations:

3.1 The set of skills drawn into a group from the curriculum standards must have a logical basis and a construct cohesiveness that both content specialists and stakeholders will recognize and support. Partitioning the standards into reporting categories must not only have a content-based logic within a subject and grade, but when feasible the schema used should be consistent across grades or levels within a given subject area.

3.2 The reporting logic should speak to the requirements of a 21st century curriculum model and address the emerging need for large-scale assessments to inform judgments about the college and career readiness of students.

3.3 The number of items or tasks drawn into a reporting category must be sufficiently large to support reliable subscores. In addition, the standards that are aggregated in arriving at subscores must be reasonably stable from one administration to the next to support consistent inferences related to performance on that reporting category. This is why reporting categories influence decisions about test length, test blueprint, and reporting.

3.4 All knowledgeable stakeholders should have the opportunity to react to the overarching logic of the reporting categories established and to the details regarding which curriculum standards are subsumed under which objectives.
Questions

- Will each item or task developed be aligned/coded to a unique Common Core standard, or will cases exist in which items are coded into multiple standards?
- Will items and tasks also be aligned/coded to other sets of standards such as those of individual states within a consortium?

Discussion – The need to document the match between a state’s curriculum standards and its assessments is partially driven by federal requirements. NCLB requires assessments to be closely aligned to standards. While the criteria for judging this alignment vary depending on the alignment study model used, the general themes are consistent: The tests must assess the breadth and depth of the curriculum standards in a balanced fashion, and the cognitive complexity level of the items should be comparable to complexity level of the standards. That is, there must be a strong match between what the curriculum standards ask students to understand and be able to do and what the tests measure. This documentation is usually achieved in part by aligning the items placed on tests to one or more standards. Cogent decisions must be made about how the items developed will be coded to the CCSS and how those alignment decisions will be reviewed and verified.

The first decision is whether items developed for a given subject and grade must be coded to a unique Common Core skill or whether items can be aligned, when appropriate, with multiple standards within the CCSS. An item or performance task could potentially assess skills from more than one standard. Although the standards typically identify discrete skills, classroom instruction rarely addresses those skills in isolation. For example, a reading lesson might have students address Vocabulary, Comprehension, and Critical Thinking skills in an integrated fashion. In a mathematics classroom, students might be found modeling a real-life situation with a function and then analyzing that function to answer a meaningful problem. Good assessment programs often seek to reflect instruction by creating more authentic assessment items or tasks that mirror the skills integration naturally found in instruction. Such items are good candidates for dual-coding.

Developing assessment items that seek to measure complex skills in isolation is possible, although somewhat artificial. When reviewing such items, committee members typically struggle to agree which one skill is being demonstrated. For example, a mathematics item that requires a student to apply the Pythagorean Theorem to find the volume of a container could be coded to a standard that deals with applying the Pythagorean Theorem, to a standard that deals with problem solving, or to both. Some of the curriculum standards established for Mathematics, Science, and Social Studies suggest the need for dual coding based on overlapping content and process skills considerations.

Twenty-first century curriculum standards are reflecting less linear and more complex and integrated curriculum models. For example, the Common Core State Standards for English Language Arts emphasize mastery of standards deemed to be “essential for college and career readiness in a 21st century, globally competitive society.” The Common Core defines both college readiness and high school standards that “work in tandem to define the college and career readiness—the
former providing broad standards, the latter providing additional specificity.” Test developers may wish to consider both when developing assessments. This could be addressed through dual-coding. While dual coding items may encourage assessing more complex skills in more authentic ways, it clearly adds a layer of complexity to test construction and score reporting.

A consortium must also determine if the items developed will be coded to one set of curriculum standards or if the items will be coded to multiple set of standards (e.g., to Common Core standards and individual state standards). Although aligning to multiple sets of standards clearly adds challenge and complexity to the process, this kind of alignment may be necessary, especially if the alternate set of standards involve different stakeholders than the primary set (e.g., college and career readiness standards), or if both national- and state-level reporting are desired.

**Recommendations:**

4.1 A consortium should make decisions about whether items will be coded to one curriculum standard or to multiple standards (within a given set of standards or across standard sets) before item development begins. The decision will influence the item development guidelines and the items, but may also affect the makeup of the test blueprint (to be discussed later) and impact scoring and reporting.

4.2 If items are to be aligned to multiple sets of standards, a consortium must make a clear decision regarding which set of standards the items will be “written to,” because the items cannot be developed with multiple masters in mind. Only one set of standards can serve as the primary curriculum alignment, and only one set can help frame the item development specifications.

5. **Content Validity**

**Questions**

- Who will review and confirm the proposed alignments of items to the Common Core standards, and what process will be used for the review and confirmation?
- How will the state or consortium document that the cognitive complexity level of the state’s curriculum standards and the state tests are aligned?

**Discussion** – As stated earlier, NCLB requires states to document the match between what the state’s curriculum standards ask students to know and be able to do, and what the tests measure.

After decisions have been made regarding the curriculum standards to which the items will be aligned, processes must be developed for making alignment decisions, reviewing the decisions, and validating the decisions. Initial alignment designations are typically made by content specialists working for item development vendors. These designations should, however, be reviewed by knowledgeable stakeholders and, ultimately, by classroom teachers. There must be a consistent process in place that serves to confirm the match between the intent of the curriculum standard and the skills the items measure.
Close alignment between the content of the items developed and the standards is best supported by the establishment of clear and specific item development guidelines, which are also called item development specifications. These guidelines are used to clarify the intent of the curriculum standards for both item writers and item reviewers.

Item development guidelines can help to define the content boundaries of the items. For example, a CCSO Mathematics standard at fourth grade may ask students to “Solve word problems involving multiplication of a fraction by a whole number.” The item development guidelines for this standard might specify limits regarding the size of the fractions to be used and could describe the types of problem contexts that are appropriate or inappropriate for that grade.

The inherent risk in specifying content boundaries is being overly prescriptive (e.g., test only fractions with denominators of 2 and 4). Such boundaries, if shared with stakeholders, could have a deleterious affect on instruction. And yet, setting boundaries can help to clarify the intended scope of instruction (e.g., specifying the assessable genres for given grade levels for Reading assessments). Policy makers must establish careful boundaries for such efforts, and the boundaries should support quality instruction and yield fair assessments.

Because the schema used to document the alignment of a consortium’s tests to the Common Core standards may also include matching cognitive complexity levels, the item development guidelines may also recommend the appropriate cognitive complexity levels for the set of items to be developed for a given curriculum standard or set of standards. This presumes that cognitive complexity assignments have been made for the curriculum standards. The items developed to a standard or to a set of standards should match, on average, the cognitive level of those standards.

A number of schemas are available for assigning cognitive complexity levels to curriculum standards and assessment items. Notable among them are the Depth of Knowledge model used in the Webb Curriculum Alignment Process \(^3\) and Bloom’s Taxonomy. Irrespective of the model used, the process used to assign cognitive complexity levels to the curriculum standards and to the assessment items should involve a representative sample of knowledgeable stakeholders and classroom educators.

**Recommendations:**

5.1 The consortia must establish and follow well-defined processes to confirm the content match between the curriculum standards and the items and to verify that the average cognitive complexity of the items matches that of the curriculum standards. These processes should include all knowledgeable stakeholders, especially classroom teachers.

5.2 Stakeholders should be careful not to define item development guidelines that are too strict. If development guidelines are too prescriptive, they may lead to “cookie-cutter” items, which could limit the success of the assessment program in reaching its stated goals. The goal is to measure student learning relative to the

---

\(^3\) Dr. Norman Webb, University of Wisconsin-Madison, Wisconsin Center of Educational Research
skills in the curriculum standards, not test-taking skills. The item development
guidelines should honor the overarching need to maintain content validity while
promoting new and innovative ways to assess the given skills.

5.3 The items developed should be reviewed and scrutinized by
knowledgeable stakeholders, including policy makers, content specialists, and
classroom educators, who are well equipped through classroom experience to help
determine if a given item appropriately assesses the curriculum standard to which it
has been aligned. Using item development guidelines alone does not necessarily
result in a close alignment between the curriculum standards and the assessment
items.

6. Item Types

Question

- Which assessment item types will be used to measure student
achievement?

Discussion – There are different ways to approach the assessment of student
achievement, including the choice of different types of items. The adoption of the
CCSS brings the promise of new and innovative assessments. Determining the
types of items to be used in these new assessments should be driven in part by
content; the item type should be well suited to assess the skills being measured. For
example, the mechanics of writing (spelling, punctuation, grammar, etc.) can be
effectively measured with multiple-choice items, but the ability to write a well
organized, well written response to an academic prompt cannot be measured
effectively with multiple-choice items. Each of the item types available has distinct
advantages and disadvantages that must be weighed.

Multiple-choice items are typically the least authentic, but the most efficient. Their
most significant drawback is that they offer students a closed universe of possible
responses, with a 25% random chance of offering the correct answer if four answer
choices are provided. Multiple-choice items can be written to varying levels of
difficulty and can address higher-order thinking skills. They work well with both
paper-based and online modalities. They are easily machine scored.

Short constructed-response items such as gridded-response and one- or two-word
response items are clearly more “open” and typically present a greater challenge for
students. They also can be written to varying levels of difficulty and can address
higher-order thinking skills. They also work well with both paper-based and online
assessments. They often can be machine scored, but doing so can present some
challenges.

Extended constructed-response items in which students must write phrases,
sentences, full-paragraph explanations, responses in mathematics items, or fully
developed compositions represent a more sophisticated assessment item type. Such
item types are especially well suited for the assessment of college and career
readiness skills. For example, student writing is regarded by the CCSSO as a key
means of showing what students know about a subject. To be college- and career-
ready writers, the Common Core standards say that students must “take their
purpose and audience into consideration … and know how to combine elements of
different kinds of writing to produce complex and nuanced writing.” Such complex
skills can only be assessed using extended constructed-response items.

Extended constructed-response items can be written to varying levels of difficulty
and can be administered using either paper-based or online assessments. Scoring
these items requires carefully developed and well-defined scoring rubrics that until
recently have been used most often by trained human scorers because of difficulties
in machine scoring these items. Recent technological breakthroughs, however, now
support analytic scoring via automated scoring engines. Extended constructed-
response items tend to be the most expensive to develop and the most expensive
and time-consuming to score; however, creative solutions are available to
overcome these obstacles, including sharing items across assessments and modified
release strategies. Please note that using such items, especially if scored by
humans, can affect scoring and reporting timelines.

Item sets must also be considered when discussing item types. Item sets are
thematically linked items using common stimuli and are most often linked to
English Language Arts assessments (e.g., students read a passage and then answer
a set of associated items all based on the passage). However, item sets are also used
in other content areas such as science. Item sets can be used in both paper and
online modes of administration. Item sets can contain only one type of item or can
be a mix of multiple item types. Item sets can include items written to varying
levels of difficulty and are especially well suited to address higher-order thinking
skills. They also work well with both paper-based and online assessments. They
can be machine scored. If utilized, item sets must be given careful consideration
when developing test designs and item development targets because they are
developed as a set and must be administered as a set.

Performance tasks are one type of item in which students engage in a series of
structured activities and then produce a product or answer a complex multi-part
question. These items are often not included in the mix of item types used in large-
scale high-stakes assessments due to time and cost constraints. However, this kind
of item offers rich opportunities to collect information about student achievement
and will be used more in future testing programs, especially online assessments.
Their use will undoubtedly offer significant new challenges for the designers of
these testing programs.

With the advent of online test administrations, the possibilities presented by more
“innovative” online item types (e.g., interactive stimuli, computer simulations)
offer great promise to create some of the most authentic item types yet. Their
development is complex, time-consuming, and expensive, which are factors that
should not be ignored. These items also can be written to varying levels of
difficulty but are especially well suited to address higher-order thinking skills.
They often can be machine scored, but doing so can present some challenges. The
use of performance tasks must also be given careful consideration when developing
test designs and item development targets because they must be administered in
their entirety.

The mix of item types used will profoundly influence the nature of the tests
administered. Item types will not only affect the “look and feel” of the test, but will
influence the perceived difficulty of the test, the cost of development, and the cost and the timelines for scoring and reporting student responses.

**Recommendation**

6.1 All policy makers and knowledgeable stakeholders must be involved in decisions about the mix of item types used in an assessment because of how significantly the mix of item types affects the overall test design, content validity of the test, item development, and the scoring and reporting of student achievement.

7. **Stimulus Types**

**Questions**
- What types of stimuli will be used in the assessments?
- Will the passage stimuli be permissioned or commissioned?

**Discussion** – The term *stimulus* can have multiple meanings in assessment. In this context stimulus refers to the content materials used to engage students in an assessment item. Examples could include the maps, charts, tables, graphs, timelines, etc. used in an individual mathematics or social studies item or in a cluster of such items, or a science simulation or scenario used to generate multiple items, or a literary selection used for a cluster of reading items.

For passage-based assessments like reading, the determination of which types of literary selections or literary genres will be used at what grade levels is critically important. For example, the Common Core Reading standards have students in K–5 applying the standards to a wide range of text types (e.g. stories, drama, poetry, and informational texts), but which genres are not stipulated in the CCSS for any particular grade. In this case the test designers may have wide latitude in selecting the genres to be used and in determining the reading level and length of the selections. Those determinations must be connected to classroom instruction and should reflect instructional practice. Classroom teachers and content specialists need to be involved in the review of all stimuli used to help determine their appropriateness for placement on the assessments.

Test designers must also determine whether the passages in an assessment will be “permissioned” (i.e., published, authentic literature that must have copyright permissions) or “commissioned” (i.e., literary selections typically authored by professional writers specifically for the assessment that adhere to detailed style and item development guidelines). Either choice has advantages and disadvantages. Permissioned pieces that are at the students’ reading level and generate rich items that align well to the curriculum standards may be difficult to find. In some cases such passages may also be expensive to use, as permissions from the copyright owners are negotiable. In general, permissioned selections cannot be edited to “fit” the items, which make them a bit harder to work with. While commissioned passages offer the advantage of being developed to “match” the test, they may not have the literary value of authentic pieces and can often be difficult to produce.

**Recommendations:**

7.1 All policy makers and knowledgeable stakeholders, especially classroom teachers, must be involved in decisions about the types of literary selections to be used in the Reading assessments.
7.2 If stimuli with copyright protections are to be used, then legally defensible permissions procedures should be established and followed scrupulously.

8. Test Blueprints for Each Subject and Grade

Questions

- For passage-based tests, what will be the number, type, and complexity of stimuli on the test?
- Which item types and how many items of each type will be on the test?
- How many points will each item/item type be worth?
- How many items/points will be in each reporting category? Will dual-coded item scores contribute to only one reporting category or both?
- If items are also coded to external standards (e.g., National Council of Teachers of Mathematic standards) will there be test construction requirements around those secondary standard sets?

Discussion – The test blueprint ultimately represents the distillation of the many decisions made about the assessment because the blueprint lays out the architecture of the test: the testing objectives, how many items/points in each objective, what type of items will be placed on the test and their relative proportions within each objective, the total number of items and score points, etc.

The test blueprint serves to articulate at a greater level of specificity many of the considerations and issues that have been discussed in this article. For example, it is one thing to determine that the type of items that will be placed on a test will be multiple-choice items, gridded-response, and extended constructed-response items. But it is another thing to quantify that and agree that the test will consist of forty multiple-choice items, five gridded-response, and two extended constructed-response items; and to then determine that each multiple-choice item will be worth one point, each griddable two points, and each constructed-response five points.

For computer-based assessments additional consideration is necessary for the inclusion of innovative item types such as hot spots, drag-and-drops, animations, and simulations. Their use in the test design is likely to increase the cognitive complexity of the test, and careful consideration must be given to their representation within the test blueprint. These decisions must be made early in the design of a new assessment program because they will also greatly affect the item development plan.

For passage-based tests, the blueprint often reflects a number of critical decisions about the type of stimuli used and the relationships between them. In a Reading test for a given grade level, for example, it may not be possible or appropriate to test all literary genres. Decisions about the literary genres to be assessed and whether there will be a connection between the literary and informational passages used will significantly influence the structure of the test. After deciding the stimuli to be used, the test designers must also specify the mix of item types that will be associated with each stimulus or across stimuli. These decisions are often reflected in the test blueprint, as well.
If items will be coded to multiple standards within the curriculum (e.g., code some items to both a content standard and a process standard) the test blueprint may also reflect quantitative decisions around such dual-coded items. For example, in a Mathematics test in which some of the items coded to specific content skills are also coded to problem solving standards, the blueprint may require that at least 30% of the items placed on the test be coded to both a content strand and a problem solving skill. The decision to dual code may allow student achievement to be reported on this test for the specific content strands and for problem solving.

As consortia of states develop large-scale assessments, the likelihood increases that items placed on such tests may be coded to multiple standard sets (e.g., to the Common Core State Standards and to individual state standards). The test blueprints, however, for such assessments would typically be developed around the primary standards (e.g., CCSS), not the secondary standards (e.g., an individual state).

**Recommendations:**

8.1 Policy makers, content experts, and the psychometrics team need to develop the test blueprint together because the test blueprint reflects content decisions and psychometric considerations.

8.2 Regardless of the format or level of granularity, the test blueprint should be officially adopted by the assessment program’s policy board because the blueprint is a policy document.

8.3 Although test blueprints for consortia assessments could reflect quantitative requirements relative to multiple sets of standards, this is not recommended because of the significant test construction challenges this would present.

9. **Psychometric Specifications**

**Questions**

- What role do psychometric specifications play in the overall test specifications?
- How do psychometric specifications relate to item development?

**Discussion** – Psychometric specifications provide guidelines for the statistical characteristics of the items or tasks that are included in the assessment, as well as the overall statistical characteristics of the assessment itself. Psychometric specifications are sometimes established at the reporting category level in addition to the overall test level. Within ongoing testing programs, psychometric specifications include additional guidelines such as the placement of items or tasks used to equate test forms from one administration to the next.

In some cases, especially when the available pool of assessment items is not large, relaxing psychometric specifications may be necessary to satisfy the test blueprint. It is not uncommon for tensions to exist between satisfying content and psychometric goals in assessment development; these tensions are best addressed through an open dialogue grounded in commonly understood principles.
Precise psychometric specifications require data that summarize the performance of items or tasks for the targeted student population. These data are typically obtained through tryout or field-test administration and include indices of difficulty and how well items and/or tasks discriminate between high-scoring and low-scoring students. Item developers need to have an idea of how difficult items or tasks are for the intended student population as early in the development process as possible. For example, a policy decision to develop a rigorous assessment may not be effectively realized if almost all of the items written for the assessment are so difficult that they can only be answered by the highest achieving students.

**Recommendations:**

9.1 Psychometric specifications should be given careful consideration by policy makers and content developers at the earliest stages of the assessment program.

9.2 Consideration should be given to piloting items through strategies such as cognitive labs early in the item development phase. Doing so can provide information to inform item writing efforts so that the statistical characteristics of the resulting items are consistent with the expectations of policy makers and psychometricians.

10. Test Administration Modality

**Questions**

- Which testing modalities (paper/online) will be used to measure student achievement?
- If online, will innovative assessment strategies be used?
- If paper, how will the use of color or other text enhancement tools (e.g., boldface or italicized text) be addressed?
- If paper and online, how will comparability be addressed?
- How will the use of student tools (e.g., dictionary, compass) be addressed?

**Discussion** – The modalities (i.e., online only, paper only, both paper and online, paper only as an accommodation) used to administer the assessment influence the student experience with the test and can significantly affect the nature of the test.

The online administration of 21st century, consortia-led assessments offers a great number of exciting and promising opportunities for new and innovative assessment strategies; but, online administration also offers some of the greatest challenges. Monographs have been written addressing these opportunities and challenges, and much of that discussion is beyond the scope of this article.

If a test will be administered in both online and paper modes, policy makers and content specialists must first determine if the online administrations will be duplicates of the paper form and differ only because they are administered via computer, or if the online administrations will be unique test builds. Even if the online tests are exact copies of the paper forms, some online issues will need to be addressed. One issue to address is how scrolling will be handled. Another issue is how tools such as protractors, compasses, rulers, dictionaries, etc., will be presented to students taking the online test. Some tools, such as a dictionary, could be provided in identical paper format or could be offered electronically online.
Tools such as a ruler or protractor would need to have online equivalents developed that met usability requirements. Although these may seem to be trivial issues, they can be challenging to resolve.

Whether the assessments are to be administered in paper or online mode, policy makers will also need to address how accessibility issues will be addressed in either or both modes. If the assessments are to be administered online, for example, will rollover audio text be provided so that students can hear certain words pronounced? Again, there are numerous publications available addressing accessibility concerns, and innovative online accessibility accommodations, and policy makers are encouraged to review them.

When a choice is made to deliver tests online, policy makers have several opportunities to consider. One opportunity is to employ the use of computerized adaptive testing (CAT) or a related use of the computer to choose and administer test content. CAT works best when items are selected from a pool one at a time but it can also work when items are associated with stimuli. CAT can result in shorter and more efficient tests, can provide immediate score reporting, and can improve test security. Other online test administration strategies besides CAT can offer similar advantages, and the decision of which strategies to employ should be based on considerations of the item types and/or stimulus types that best measure the curriculum standards.

Online testing also may include the use of technology-based, innovative item types. Innovative item types can offer rich ways to assess skills that are often difficult to address in paper mode. Online innovative items can include videos and animations, can provide “shared experiences” or allow explorations using virtual technologies. Please note that although innovative online item types offer exciting new opportunities, their use also presents unique challenges. These items may be more difficult and expensive to develop, may require more administration time, and may present scoring challenges.

Numerous issues associated with the paper administration of assessments must also be addressed. Most of these issues are issues in the online world as well and will need to be resolved in either or both administration modes. Most policy makers and content specialists are familiar with these issues and are well equipped to resolve them. Test development issues such as font; item and test layout; permissible strategies for text emphasis; the use of color; style conventions for charts, tables, graph, and other graphics; etc. must all be decided before item and test production can begin.

When assessments are delivered in both paper and online modalities, some immediate test construction challenges will need to be addressed. Many online items, especially innovative online items, cannot be delivered comparably in paper mode; therefore, comparable tests cannot be built in terms of content, skills and psychometric targets. Comparability studies will be required that may include an assessment of difficulty and psychometric test structure. Based on comparability study results, the online and paper modes may be judged to be comparable or in some cases, score conversions are adjusted to establish comparability between modes. But comparability is also perception. A paper test without innovative
dynamic online items, even though psychometrically equivalent and addressing the same content, may not be perceived by the stakeholders as a comparable test.

**Recommendations:**

10.1 If tests are to be offered in both online and paper modes, policy makers and stakeholders will need to address practical test administration issues (e.g., the use of tools) before item development begins.

10.2 If tests are to be offered in online mode, policy makers must decide if innovative online item types will be included (e.g., computer simulations, “drag-and-drop” items, the use of video animations as stimuli, dynamic graphing options in mathematics, etc.).

10.3 The resolution of “style” issues for paper and/or online administrations should be handled through the development of a “style guide.” The style guide should be adopted by the assessment program’s policy board. All stakeholders who will review items should adhere to the style rules.

10.4 If tests are to be offered in both online and paper modes, policy makers will need to address both the psychometric comparability issues and the perception issues.

11. **Performance Scoring Models**

**Questions**

- How will constructed-response (short or long) items be scored?
- Will a holistic or analytic scoring model be used to score student writing samples?
- Will student writing samples be scored by human scorers or by computer scoring engines?
- What scoring agreement model will be used?
- How will rangefinding be conducted?
- Who will monitor the performance scoring program?
- What issues must be addressed for multiple-choice, gridded-response, and hot-spot items that are to be scored by machine?

**Discussion** – Standardizing the scoring of responses to performance tasks, such as constructed-response items, is crucial to producing reliable and valid assessment results. A human element is often involved in scoring these responses, and standardizing human judgment can be a complex matter. While automated scoring routines recently have been used to help reduce the cost of scoring, such routines, while standardized, may offer pedagogical and perceptual challenges. Ultimately, the philosophical and financial considerations of the policy makers will determine which scoring model will be selected for either a short or long constructed-response items.

Policy makers must also determine whether student responses should be scored using a holistic (i.e., as a single body of knowledge), an analytic (i.e., with discrete skill checkpoints), or a combined model. For example, an analytic model requires scorers to determine to what degree disparate features of a response are present (or absent), but a holistic model requires that scorers use a rubric to guide them into
making a single evaluation of the response as a whole. Still other models (i.e., trait scoring and core scoring) combine elements of the holistic and analytic models to fit the needs of a given assessment. The most important factor in determining a scoring model is the purpose of the assessment.

Once a scoring model has been determined, policy makers must weigh factors such as the number of score categories for each task, the specific criteria for scoring each task, the number of scorers who will evaluate each response, the agreement model used by readers to determine student scores, and the monitoring of scorers to achieve validity and reliability. While these factors are often dependent upon financial constraints, they are inherently philosophical as well. The number of score categories for each task directly affects the way in which the response is made manageable for scorers. Furthermore, to determine specific criteria for scoring each task, purposeful rubric design and/or skill checklists are essential and must be tied to the mode of response being assessed and the scoring model selected. The type of agreement method chosen to be employed in scoring these responses—perfect or adjacent—also becomes important to the reliability and validity of the assessment results. Finally, the training and monitoring of scorers depends upon the way in which ranges within and across score points, often called rangefinding, have been established and the method used to validate that they are being actualized.

Machine scoring of items also requires consideration in some cases. Traditional multiple-choice items do not require any additional input or adjudication during scoring. However, machine-scored items such as gridded-response or innovative items such as hot spots or drag-and-drops require scoring rules to be defined in order to be accurately scored. Ideally, the scoring rules are defined in advance of the start of development. Items are developed according to the specifications and all possible correct answers or correct locations are collected and verified for scoring during the item development. Once scoring has begun, an additional check verifies the accuracy of the scores that have been previously specified.

The perceived benefits of constructed-response items must be weighed carefully when designing an assessment. While certain limitations in cost, time, and complexity are placed upon constructed-response items, they remain a valid means for assessing student populations in a way that multiple-choice items cannot achieve. Ensuring that the design and scoring of these items are as standardized as possible is crucial. Test reliability and scoring quality can be problematic with constructed-response items when deliberate item and rubric construction, comprehensive scorer training and evaluation, and consistent monitoring of scoring quality are not completed. However, when used appropriately and thoughtfully, constructed-response items can provide insight into student achievement across complex performance skills.

Recommendations:

11.1 The scoring model used and the specific details surrounding the scoring rubric must be resolved before the item development specifications and the constructed-response items are developed.

11.2 Policy makers should determine the number of score categories for each performance task, the specific criteria for scoring each task, the number of scorers
who will evaluate each response, the agreement model to be used by readers to
determine student scores, and the monitoring of scorers that will be used to achieve
validity and reliability before the constructed-response items are developed.

11.3 If using certain machine-scored items such as gridded items, hot spots,
and drag-and-drops, additional scoring considerations that must be addressed
before item development begins.

12. Accessibility Guidelines

Questions

- How will universal design (UD) principles be best incorporated in item and
test development requirements or item evaluation criteria?
- How will items be reviewed or evaluated relative to accessibility
    standards?

Discussion – Federal guidelines have increasingly required assessment systems to include
students with disabilities and those with limited English proficiency. Rather than retrofitting
assessments with accommodations that are only partially effective and/or compromise
validity, or developing a variety of alternate assessments to include these students, the
proposed new 21st century assessments should be designed and developed to permit inclusive,
fair, and accurate testing of the widest possible range of students. This should be done in a
way that results in valid inferences about performance for all students who participate in the
assessment.1 The principles that guide such development are generally called universal design
principles. A universally designed assessment will anticipate the variety of accessibility needs
of potential students and build in methods that allow all students to access, engage with, and
respond to test content in the most accessible manner possible.2 As emphasized by the
National Research Council in their Committee on Appropriate Test Use report,

“Fairness, like validity, cannot be properly addressed as an afterthought once the
test has been developed, administered, and used. It must be confronted
throughout the interconnected phases of the testing process, from test design and
development to administration, scoring, interpretation, and use.”3

Tests should be designed to minimize potential sources of construct-irrelevant
variance by supporting the ways that diverse students interact with the assessment
process.4 For this reason, state policy makers and content specialists should establish
guidance regarding the accessibility of items and tests for special populations early in
the planning process. Such guidance is intended to help item writers and reviewers
develop items that are accessible to diverse student populations so that performance
for all students can be measured as accurately as possible.

Specifically, UD principles address issues related to the reading level and contextual
information contained in items and stimuli. They typically provide suggestions
regarding language structure and syntax, vocabulary, accessible and inaccessible
contexts, the use of graphics, and other devices that do not affect the construct being
measured, but increase students accessibility to the assessment and support
administration accommodations. Universal design guidelines also address modes of
test presentation and response, to maximize students’ opportunities to demonstrate
their construct-relevant knowledge, skills, and abilities5.
The criteria used to review items and passages during the development process should specifically include review criteria relative to UD principles. Policy makers, content specialists, and educators who participate in the review process should all be provided training in universal design and be sensitized to the accessibility issues associated with UD. Conducting the accessibility review of items separate from the content review might be advisable to encourage participants to focus on accessibility issues. Bias reviews are often conducted incorporating UD principles. And finally, accessibility reviews should address test interfaces, not just test content.

While research has demonstrated that principles of UD can be used to make a more inclusive testing environment that reduces the need for accommodations for special populations, it should be noted the use of universal design principles increases the accessibility of the assessment for all students, and is supported by almost all professional assessment organizations. And finally, as we move toward the development of truly new, 21st century assessments, it should be recognized that UD principles can also guide the development and delivery of “innovative items,” those that use digital technologies to test students on greater depths of knowledge and skill than traditional items. Again, there are various publications that offer specific universal design guidelines that can be used as the basis for establishing accessibility principles to guide these new assessment programs.

Recommendations:

12.1 Policy makers and content specialists should work collaboratively with all knowledgeable stakeholders, especially classroom educators and special-needs experts, in prioritizing the order in which UD principles are applied during the design and development of both the test content and testing interfaces used in the assessment program.

12.2 The UD principles agreed upon should be adopted by the assessment program’s policy board. The principles should be shared with the stakeholders who will review items, and UD training should be provided for all participants who will review and approve items.

12.3 State policy makers should establish guidelines for the accessibility of items and tests for special populations before item development begins.

References:


13. Item and Stimulus Review

**Questions**

- What criteria will be used to review items and stimuli?
- How will review meetings be conducted?
- Who will participate in such meetings, and how will participants be selected?
- At which meetings will edits be permitted, and how will the final set of edits be implemented be established?

**Discussion** – The life cycle of an item or passage on a large-scale high-stakes assessment is typically two or more years. During that time the items are reviewed by numerous individuals and groups and are edited and revised repeatedly. Ultimately, the items are field tested, student performance data is collected and analyzed, and the items together with their data are again reviewed for their fitness to be placed on tests. Items found suitable for use on tests are typically placed in a statistical item bank, and tests are eventually built using items from the bank. Those tests may be subjected to additional review by individuals and/or groups before their administration.

Item review guidelines typically list the specific criteria individuals or groups reviewing items should use in judging the appropriateness of the items. Review criteria should be developed that address the content of the item (e.g., is the item clear, is the content accurate, does the item address the curriculum standard to which it is aligned, is the correct answer correct, is there only one correct answer, are the distractors plausible, are the distractors unequivocally wrong, etc.). Review criteria may also ask individuals to make judgments about other item attributes such as readability, reading load, grammar, spelling and punctuation, cognitive complexity, accessibility, cultural bias, etc. Content-area specific review criteria may include issues such as grade-appropriate vocabulary usage; format for tables, charts, equations, and graphs in mathematics; or format requirements for maps in Social Studies items; etc. Similar criteria should be developed for other stimuli. The list of review criteria could be quite extensive.
After review criteria have been developed, the process by which items will be reviewed should also be clearly articulated. Typically, multiple review stages will involve different professionals at each stage. The policy makers should agree on the review stages and who will participate in each review. If external reviewers are to participate (e.g., K-12 educators, university professors), clear guidelines should also be established regarding their selection and the conduct of such meetings. If training modules are to be developed for these external review meetings, the assessment program policy makers and content specialists should agree on the content and delivery of such training.

Clear guidelines also are needed to govern who can make edits and when edits can be made because items are typically edited frequently. In addition, a clear decision is necessary to indicate specifically who determines which edits will be implemented. Given the tendency of reviewers to want to continuously make edits, agreement is needed about who signs off on the final edited version as well as a very clear agreement that no further edits can be made to the item after that stage.

Recommendations:

13.1 Whether the review criteria for an assessment program are developed at a high level or granular level, they should be established early in the item and stimulus development process, and the final versions should be shared with all individuals and groups who will be conducting such reviews.

13.2 The review criteria should be developed with input from all stakeholders and may be adopted by the assessment program’s policy board.

Conclusion

The Common Core curriculum standards will undoubtedly help to transform the educational landscape over the next decade and beyond and will clearly serve as guidelines for instruction in participating states. Comprehensive test specifications are needed to translate those common standards into a sound and effective 21st century assessment program for the consortia.

Test specifications serve to delineate the purpose of the assessment, the underlying architecture of the tests, and the content boundaries of the assessment program. A clearly articulated set of test specifications contribute important validity evidence for the assessment and provide the basis for access, fairness, and reliable scores. Test specifications should be established before item development begins. We hope that the recommendations offered in this article will support that development effort. Policy makers and stakeholders are encouraged to address the issues raised in this article when formulating those specifications. Through thoughtful deliberation and careful planning by the consortia around these test specification issues, we believe a high-quality 21st century assessment program will emerge.