Smarter Balanced Assessment Consortium

General Item Specifications

Developed by Measured Progress/ETS Collaborative
April 16, 2012

Smarter Balanced General Item Specifications

Table of Contents

Smarter Balanced Assessment Consortium: General Item Specifications DRAFT

Introduction

This document and its companion documents are designed to provide item developers specific guidance in creating items that meet the expectations of the Common Core State Standards according to the criteria outlined by Smarter Balanced. It is a highly detailed document that outlines complex requirements for a very specific audience: the item and task developers and reviewers. Table 1 provides a list of the overall set of materials that are part of the Item Specifications work for Smarter Balanced. This general document will provide summary information about each companion document.

Table 1. Smarter Balanced Assessment Consortium Item Specifications and Guidelines Documentation.

Deliverable	Brief Description
General Item Specifications	This document, which provides an overview of the body of work as well as detailed information for topics not covered in other documentation. This document covers both Mathematics and English Language Arts and Literacy. In addition to this document, there is a general specifications document for each content area that covers guidelines for item writers specific to the content area.
Mathematics Item Specifications	The Mathematics Item Specifications include guidance specific to writing Smarter Balanced Mathematics Items and a table for each claim and target combination expected to be addressed by the summative assessment. The item specification tables list the evidence required to establish the claim and target, and include detail on DOK level, CCSS standards covered, a list of item models matched to the evidence statements and other information pertinent to item design,
English Language Arts and Literacy Item Specifications	The ELA Item Specifications include guidance specific to writing Smarter Balanced ELA Items and a table for each claim and target combination expected to be addressed by the summative assessments. The item specification tables list the evidence required to establish the claim and target, and include detail on DOK level, CCSS standards covered, a list of item models matched to the evidence statements and other information

	pertinent to item design.
Sample Items	For both mathematics and ELA there are sample items associated with selected item models within the claims and targets.
Style Guide	The Style Guide provides style conventions and specifications for both test items and test forms created for the Smarter Balanced Assessment Consortium summative assessments.
Stimulus Specifications	The Stimulus Specifications provide guidance to item writers regarding selection and evaluation of stimulus material for ELA assessments. Mathematics stimulus guidance can be found in the Mathematics General Specifications
Technology-Enhanced Item Specifications and Templates	The Technology-Enhanced component of this work includes general specifications for the development of TE items, as well as 25 templates and 35 sample items.
Performance Task Specifications	The Performance Task Specifications provide general guidelines for the development of performance tasks.
Bias and Sensitivity Guidelines	The *Bias and Sensitivity Guidelines* will help ensure that the Smarter Balanced assessments are fair for all groups of test takers by providing guidelines for item and task writers as well as reviewers to follow as they evaluate suitability for content in assessments.
Accessibility and Accommodations Guidelines	The Accessibility and Accommodations Guidelines include six documents that are intended to be used by item writers and accessibility experts to make items and tasks accessible to as many students as possible.

An Introduction to the Smarter Balanced Assessment Consortium

The Smarter Balanced Assessment Consortium is one of two multistate consortia awarded funding from the U.S. Department of Education to develop an assessment system based on the new Common Core State Standards (CCSS). To achieve the goal that all students leave high school ready for college and career, Smarter Balanced is committed to ensuring that assessment and instruction embody the CCSS and that all students, regardless of disability, language, or subgroup status, have the opportunity to learn this valued content and to show what they know and can do. With strong support from participating states, institutions of higher education, and industry, Smarter Balanced will develop a balanced set of measures and tools, each designed to serve specific purposes. Together, these components will provide student data throughout the academic year that will inform instruction, guide interventions, help target professional development, and ensure an accurate measure of each student's progress toward career- and college-readiness.

Core Components of Smarter Balanced

Summative assessments:

- Mandatory comprehensive accountability measures that include computer adaptive assessments and performance tasks, administered in the last 12 weeks of the school year in grades 3–8 and 11 for English language arts(ELA)/literacy and mathematics;
- Designed to provide valid, reliable, and fair measures of students' progress toward and attainment of the knowledge and skills required to be college- and career-ready;
- Capitalize on the strengths of computer adaptive testing (e.g. efficient and precise measurement across the full range of achievement and quick turnaround of results); and,
- Produce composite content area scores, based on the computer adaptive items and performance tasks.

Interim assessments:

- Optional comprehensive and content-cluster measures that include computer adaptive assessments and performance tasks, administered at locally determined intervals throughout the school year;
- Results reported on the same scale as the summative assessment to provide information about how students are progressing;
- Serve as the source for interpretive guides that use publicly released items and tasks;
- Grounded in cognitive development theory about how learning progresses across grades and how college- and career-readiness emerge over time;
- Involve a large teacher role in developing and scoring constructed response items and performance tasks;
- Afford teachers and administrators the flexibility to:
    - select item sets that provide deep, focused measurement of specific content clusters embedded in the CCSS;
    - administer these assessments at strategic points in the instructional year;
    - use results to better understand students' strengths and limitations in relation to the standards;
    - support state-level accountability systems using end-of-course assessments.

Formative tools and processes:

- Provides resources for teachers on how to collect and use information about student success in acquisition of the CCSS;
- Will be used by teachers throughout the year to better understand a student's learning needs, check for misconceptions, and/or to provide evidence of progress toward learning goals.

The Role of Evidence-Centered Design

Evidence-centered assessment design (ECD) is an approach to creating educational assessments in terms of evidentiary arguments built upon intended constructs, with explicit attention paid to the potential influence of unintended constructs (Mislevy, Steinberg, & Almond, 2003). ECD accomplishes this in two ways. The first is by incorporating an overarching conception of assessment as an argument from imperfect evidence. This argument makes explicit the claims (the inferences that one intends to make based on scores) and the nature of the evidence that supports those claims (Hansen & Mislevy, 2008; Mislevy & Haertel, 2006). The second is by distinguishing the activities and structures involved in the assessment enterprise, in order to exemplify an assessment argument in operational processes. By making the underlying evidentiary argument more explicit, the framework makes operational elements more amenable to examination, sharing, and refinement. Making the argument more explicit also helps designers meet diverse assessment needs caused by changing technological, social, and legal environments (Hansen & Mislevy, 2008; Zhang et al., 2009).

The ECD process involves five layers of activities. The layers focus in turn on the identification of the substantive domain to be assessed; the assessment argument; the structure of assessment elements such as tasks, rubrics, and psychometric models; the implementation of these elements; and the way they function in an operational assessment, as described below.

Domain Analysis. In this first layer, domain analysis involves determining the specific content to be included in the assessment. For Smarter Balanced, domain analysis was conducted by the developers of the Common Core State Standards which define the domain to be assessed by the Smarter Balanced Assessment System.

Domain Modeling. In domain modeling, a high-level description of the overall components of the assessment is created and documented. For Smarter Balanced, the components of the assessment system were articulated in the proposal to the Race to the Top Assessment Program. At a high-level, the components include computer-adaptive summative assessments in mathematics and ELA, interim assessments, and materials that support formative assessment practices.

The Conceptual Assessment Framework. Next, the conceptual assessment framework is developed. In this layer, the knowledge, skills, and abilities to be assessed (otherwise referred to as the *intended constructs* or the *targets of assessment*), the evidence that needs to be collected, and the features of the tasks that will elicit the evidence are specified in detail. Ancillary constructs that may be required to respond correctly to an assessment task but are not the intended target of the assessment are also specified (for example, reading skills in a mathematics examination). By identifying these ancillary KSAs, construct-irrelevant variance can be identified up-front and minimized during item and task development—potential barriers created by the ancillary KSAs can be removed or their effects reduced through the provision of appropriate access features. For Smarter Balanced, the constructs that are the target of assessment are defined in the Content Specifications. Ancillary constructs are elaborated on in the Item Specification Tables. The evidence required to support claims about the assessment targets is also defined in the Item Specification Tables.

Implementation. This layer involves the development of the assessment items or tasks using the specifications created in the conceptual assessment framework just described. In addition, scoring rubrics are created and the scoring process is specified. For Smarter Balanced, items, performance tasks, and associated scoring rubrics will be developed starting in the spring of 2012 by educators and contractors.

Delivery. In this layer the processes for the assessment administration and reporting are created.

Throughout the development of item specifications, principles of Evidence-Centered Design were employed in four ways. For Smarter Balanced, the administration and reporting procedures will be developed through future contracted work.

- First, claims intended to be made based on assessment results and the constructs that are the focus of those claims were carefully considered. These claims and targets of assessment (a.k.a. assessment targets) were intended to be developed and fully articulated prior to the development of item specifications. However, challenges articulating these claims and assessment targets necessitated co-development of assessment targets and the item specifications. This provided an opportunity to refine the assessment targets to assure they were assessable in the context of Smarter Balances assessment design.
- Second, evidence required to support claims about a given assessment target or set of targets was specified. The evidence required took the form of statements about the products students are expected to produce that serves as evidence for a specific claim or set of claims about an assessment target or set of targets. The evidence required defines the information to be provided by a given item or task.
- Third, task models were designed to elicit the required evidence for a given claim or set of claims about an assessment target or set of targets. The task model provides a description of key features of items and tasks that may be developed from the model. Among the key features specified are the general contents of an item prompt, characteristics of accompanying stimuli, the type of interaction the student is expected to perform with item or task content, the expected response type, and, for selected responses, characteristics of response options.
- Fourth, two additional pieces of information were specified to help clarify aspects of the construct that were essential and non-targeted constructs that may create unique challenges for the measure of the assessment target. Specifically, domain specific vocabulary and prior knowledge that is considered an aspect of the targeted construct was specified. This information clarifies terminology and skills that are essential for valid measure of the assessment target. This information is useful when considering accessibility considerations and indicates content and skills that should not be compensated for through accessibility supports. Identification of non-targeted constructs that may create unique challenges is also useful for informing efforts to improve the accessibility of an item or task. Specifically, this information identifies unique challenges associated with the item model. As an example, an item model that requires students to produce detailed or accurate geometric figures may create unique challenges for students with fine motor skill challenges. Collectively, this fourth aspect is essential for informing efforts to improve the accessibility of items and tasks, and helps clarify the construct(s) that is the target of assessment.

In addition to the above, principles of Evidence-Centered Design were also applied to develop a sample of items for a sub-set of task models. These sample items and tasks were developed to exemplify how the Evidence- Centered Design task models can be used to guide the development of items and tasks that elicit the required evidence used to support one or more claims about one or more assessment targets.

Alignment Framework

Background

In developing a system of assessments, Smarter Balanced is committed to ensuring that its measurement reflects the expectations of content, rigor, and performance that make up the Common Core State Standards. To that end, Smarter Balanced has designed its item specifications to demonstrate alignment through alignment methodologies reflective of Evidence-Centered Design theory. That alignment begins with an understanding of the goals of aligning assessments and standards – especially in light of the ECD approach used by Smarter Balanced.

According to Norman Webb (2002), "alignment of expectations for student learning and assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system." DeMauro (2004) states, "Alignment activities…should be the guiding principal of test design, and item alignment studies should be sources of validity documentation, as should any studies of test content." Clearly, there is a close connection between validity and alignment, validity addressing the appropriateness of inferences drawn from test results and alignment having to do with "how well all policy elements [e.g., expectations and assessments] guide instruction and, ultimately, student learning (Webb, 1997).

The critical nature of content alignment became clear to all educators as a result of the Debra P. vs. Turlington case in 1981, in which it was ruled that the content of a test must be aligned to curriculum/instruction to be fair. This is intended to be accomplished by both being aligned to the same content standards, thereby assuring that students have had the opportunity to learn the tested material. Indeed, ESEA now requires that state accountability assessments be aligned with state content standards. With most states having adopted the Common Core State Standards in English Language Arts/Literacy and in Mathematics, and with the Smarter Balanced Assessment Consortium funded to develop the next generation assessments for the consortium states, it is imperative that Smarter Balanced conduct appropriate alignment studies. Webb (1997) identifies several categories of criteria for judging alignment. This section addresses the one that is most relevant to the activity of developing items: content focus – specifically, how well the Smarter Balanced tests and items/tasks will address the expectations embodied in the Smarter Balanced Content Specifications and the Common Core State Standards.

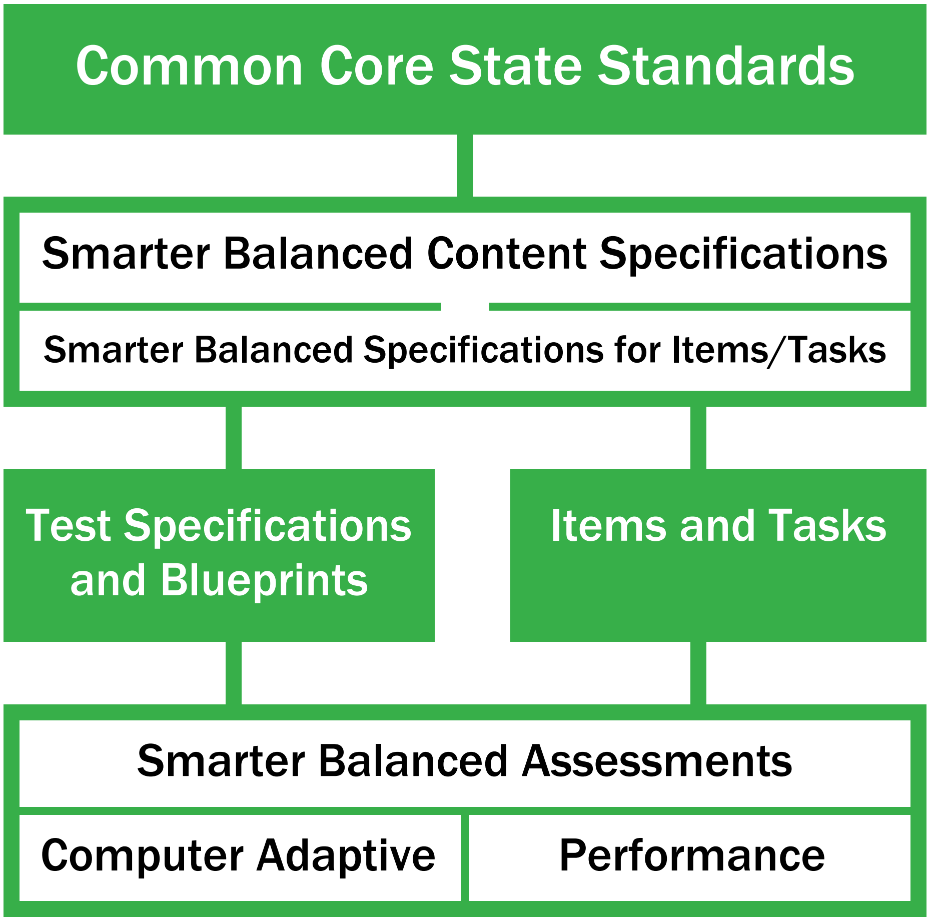
Test content alignment is at the core of content validity and consequential validity (Martone and Sireci, 2009). Because of the high stakes associated with statewide testing during the NCLB era, more attention than ever before has been given to test alignment. The emphasis on test content in alignment and validity studies is understandable. After all, a test is a small sampling of items from a much larger universe of possible items covering, at least in state assessments, a very broad domain. Thus, for inferences from test results to be justifiable, that sample of items has to be a good one – a good representation of the broad domain, providing strong evidence to support claims based on the test results. The sections that follow will explain the structure of Smarter Balanced content and discuss alignments within pairs of elements in that structure.

Smarter Balanced Content Structure

Typically, discussions of content alignment address the direct relationships between items and standards and between tests (collections of items) and standards. However, the Smarter Balanced development approach, which makes use of Evidence- Centered Design, has created levels of test and items specifications that necessitate a validity argument in the form of a chain of reasoning (See

Figure 1). Described below are the elements of the Smarter Balanced content structure and the linkages among them for which alignment studies could be considered. However, given the purpose of test alignment studies, all possible linkages do not need to be considered. Ultimately, two alignment studies are planned, along with a standards validation study.

Figure 1. Smarter Balanced Content Structure.


Common Core State Standards
Smarter Balanced Content Specifications
Smarter Balanced Specifications for Items/Tasks
Test Specifications and Blueprints
Items and Tasks
Smarter Balanced Assessments
Computer Adaptive
Performance

Common Core State Standards (CCSS). These are the content standards in English Language Arts/Literacy and Mathematics that most states have recently adopted.

Smarter Balanced Content Specifications in English Language Arts/Literacy and Mathematics. The CCSS were not specifically developed for assessment and contain a great deal of rationale and information about instruction. Therefore, following a practice many states have used in the past, Smarter Balanced distilled from the CCSS a set of content specifications expressly created to guide assessment development. Within each of the two subject areas at grades 3 through 8 and high school, there are four broad claims, and within each claim there are several assessment targets. The claims in ELA/literacy pertain to reading, writing, listening/speaking, and research. In mathematics, the claims pertain to concepts/processes, problem solving, communicating reasoning, and modeling/data analysis. Because of the breadth of the individual claims, the targets within them really define the nature of the performance expectations within these statements.

Smarter Balanced Specifications for Items and Tasks. These specifications are extensions of the Smarter Balanced content specifications. For every target, a table has been produced describing (1) evidence to be gathered to address the target and (2) several models for items to be developed to measure student performance relative to the target.

Smarter Balanced Items and Tasks. The item/task specifications and sample items developed from them are intended to guide item and tasks developers in the future. The item/task types are selected-response, constructed-response, extended-response, technology-enhanced items, and performance tasks.

Smarter Balanced Test Blueprint. Test specifications (including a blueprint) will describe the make-up of the two assessment components (computer adaptive test and performance assessment) and how their results will be combined and reported.

Smarter Balanced Assessment Instruments. These are the tests that will actually be administered to students. For the computer adaptive component, the specific items administered to students are uniquely determined for each student based on the item-selection algorithm to be developed for the adaptive testing system. The performance tasks will be administered on a matrix-sampling basis. Thus, test alignment studies will not simply be studies of a limited number of fixed tests.

Alignment Considerations

Before identifying the linkages between content elements to be evaluated by alignment studies, some mention of criteria to be considered in the studies is warranted. While there are other models whose criteria could be considered for the studies, some detail on the criteria in the Webb Model is provided here because this model is the one most commonly used for state assessments and conceptually it may be the most basic. Furthermore, other models actually use Webb's criteria among others.

Webb alignment criteria. Webb identifies four primary types of content alignment. He defines others, but the four described below are the focus of alignment studies. (Webb uses "standards" to refer to the broadest categories of content within a subject area, and "objectives" to mean the next level of content/skill categories within standards. The Smarter Balanced counterparts of standards and objectives are claims and targets.)

<u>Categorical concurrence</u> refers to the commonality between the content categories of the standards and the content categories of the assessment [items]. However, in assessment alignment studies it means more than this: it refers to the extent to which items in a test can be matched to objectives within the standards. ***<u>Depth of knowledge (DOK) consistency</u>*** between standards and assessment refers to the match between the cognitive demand of items and the level of cognitive demand communicated by the wording of the objectives. ***<u>Range-of-knowledge</u>*** pertains to the number of objectives within each standard that are covered by an item or items. ***<u>Balance-of-representation</u>*** addresses the relative coverage of content categories (standards or objectives within standards) by items in a test – i.e., the degree to which one standard or objective is given more emphasis in the assessment than another. When applied in a study of the alignment between content standards and an assessment, the Webb model provides quantitative measures and criteria for the four alignment types.

Smarter Balanced alignments. Table 2 identifies six pairings of content elements that should be aligned. Each pairing is then discussed in the context of whether or not a separate alignment study is necessary. However, the two studies ultimately planned directly address the critical alignment between the actual tests students take and content standards/specifications.

Table 2. Content Pairings.

Possible Alignments	Criteria Involved	Alignment Study Recommendation
1. Content specs (targets/claims) to CCSS	Categorical concurrence, depth of knowledge	No
2. Item models to targets (content specs)	Categorical concurrence, depth of knowledge	No
3. Items/tasks (sample) to models/targets	Categorical concurrence, depth of knowledge	Yes
4. Test blueprint to content specs	Balance of representation (content & DOK), range of knowledge	No
5. Tests administered to students to blueprint	Balance of representation, range of knowledge (categorical concurrence & depth of knowledge assumed based on Study 3)	Yes
6. Student response sets to performance level descriptors	Standards validation (judgmental or empirical verification of descriptors)	Yes

Note: All item types will be considered in alignment studies. A single performance task includes measures of different claims and targets. For purposes of alignment studies and reporting with respect to claims, individual measures within a performance task will be evaluated within their respective claims and targets.

Content specifications (targets/claims) to CCSS. It is a logical and appropriate assumption that these two documents are well aligned. As explained earlier, the Smarter Balanced content specifications in ELA and mathematics are a distillation of information from the Common Core State Standards. In fact, every target in the content specifications is mapped to the corresponding objective(s) in the CCSS. Furthermore, significant attention was given to consistently identifying DOK levels for the targets. In the development of the Smarter Balanced item and task specifications, the content specifications alignment was considered *a priori*.

Item/task models to targets (content specifications). Often in statewide assessment, item writers are given the objectives and asked to develop a certain number of items of different types and at different DOK levels to address the objectives. In the case of Smarter Balanced, item/task models have been developed for each target to assure a greater degree of consistency and replicability of the items or tasks addressing a target across developers and across years. Rather than separately evaluating the alignment of models to targets, however, it would seem that the most important test of the models would be whether the items and tasks they yield align well with the targets (the content specifications). Thus, the study described below will serve to evaluate this alignment. If many items/tasks submitted for a particular model tend not to fit the targets, then Smarter Balanced will consider revising the model.

Items/tasks (sample) to models/targets. The study planned below is intended to verify that the items/tasks written from models align with the models from which they were generated and with the corresponding targets. Thus, this study will be conducted early in the item development process so that adjustments to a model can be made should it be determined that a misalignment between items and a target is the result of a model and not the quality of item writing. The relevant alignment criteria for this study will be categorical concurrence and depth of knowledge.

Another type of study focused on items is planned, given the intent of Smarter Balanced to use new item types, particularly technology-enhanced items (TEIs). To determine whether items are truly measuring intended competencies, Smarter Balanced has planned to use *cognitive labs* . Cognitive labs use verbal reporting and think-aloud protocols in conjunction with interviews of students to identify the mental processes students use when completing tasks (Zucker, Sassman, and Case, 2004). The qualitative information they yield to complement quantitative data will be particularly useful in validating new item types and measures of more complex reasoning and performance. In fact, using the cognitive lab approach for both the new TEIs and traditional items addressing the same knowledge and skills will likely allow comparisons across those item types to identify the extent to which they are measuring the same or different competencies.

Test blueprint to content specs. The test blueprint(s) for Smarter Balanced will be the product of the test designers – the Smarter Balanced curriculum experts and leadership, with recommendations from the external measurement experts. How a blueprint will reflect the content specifications in terms of numbers of items of different types and their distribution across content categories and DOK levels will be the result of human judgments by these individuals. Thus, in many ways the relationship between a blueprint and the content specifications is more a matter of policy informed by content and measurement expertise, and less a matter of direct alignment.

The close association between a test blueprint and the adaptive test algorithm should be mentioned here. How well students' tests align with a blueprint, to be evaluated as described below, will be largely a function of the effectiveness of the algorithm that has yet to be developed.

Tests administered to students to blueprint. Frequently, a standards/assessment alignment study addresses all four of Webb's types of alignment. However, for many high stakes assessment programs, so much expert judgment goes into the proper assignment of items to content categories and cognitive complexity levels, that the individual item alignments (categorical concurrence and depth of knowledge) for items selected for final test instruments may not have to be studied. For Smarter Balanced, those alignment criteria are addressed in the item/task to models/targets study discussed above. This planned study is intended to verify the extent to which the combined adaptive and performance tests that are administered to individual students meet the specifications defined by the test blueprint in terms of balance of representation and range of knowledge.

Student response sets to performance level descriptors. Twenty years ago, as large scale accountability assessment became more prevalent, standard setting methods came under greater scrutiny. In 1992, the National Academy of Education sponsored an evaluation of performance standards established for the National Assessment of Educational Progress. One concern raised as a result of the study was an inconsistency between what performance level descriptors said students were able to do and what they were actually able to do based on further investigation. Given the secure nature of adaptive tests and the fact that the bulk of the Smarter Balanced assessments will be easily machine scored, and depending on the standard-setting method used for Smarter Balanced, Smarter Balanced intends to validate the performance standards (cut scores).

Recommended Studies

Items/tasks to models/targets. For each sampled model, panelists will be asked to judge whether (1) the items measure the target intended by the model and (2) whether the items are written at the DOK level targeted by the model. For each model, the percentage of items aligned to (1) the target,

(2) the appropriate DOK, and (3) both criteria, will be computed. These percentages will then be averaged across models (and the range computed) to give an overall indication of item-to-model alignment. Corrective action regarding the clarity and fit of a model to its target will be taken if necessary. (Note: A particular model may identify multiple DOK levels. DOK concerns usually pertain to the under representation of items at the higher levels. It will be important that items sampled for any particular model represent all the DOK levels identified by the models.)

Tests administered to students to blueprint. For each sampled student, the items and tasks comprising his or her test "form" will be compared to the test blueprint, and appropriate alignment statistics will be computed. For range of knowledge, Webb requires that for a test to be considered aligned, 50 percent of the objectives (targets) within every standard (claim) must be measured by at least one item. For balance of representation, the Webb model yields an index that is a measure of how evenly items are distributed across objectives measured within standards. Balance of representation will be examined for both content claims/targets and depth of knowledge levels. Representation relative to assigned proportions in the blueprints rather than equality of representation will be the focus of Smarter Balanced.

Determining the most appropriate quantitative criteria for Smarter Balanced will require further deliberation. While the Webb criteria should be considered seriously, it may be that other statistics will be more appropriate, given the number of targets and other unique features of the Smarter Balanced system. It is likely that the statistics computed for the individual test forms will be aggregated across sampled forms. For example, if Webb criteria were used, the percent of forms meeting the 50 percent rule for range of knowledge and the average balance of representation index would be computed. Ideally, students should, on average, receive test forms that are highly aligned with the test blueprint.

Student response sets to performance level descriptors. This recommended study is not a typical alignment study applying Webb-like alignment criteria. It is a study to see if what a performance level descriptor says students at a particular proficiency level are able to do matches what they can actually do. This may be done through either (1) a standards validation process involving panels of judges matching whole bodies of student work on Smarter Balanced assessments to performance level definitions or (2) a validation study involving other measures of student competencies that would serve as acceptable, authentic criterion measures.

Conclusion

Clearly, the alignment between expectations of students and Smarter Balanced assessments is essential. This paper discussed alignments between components of the Smarter Balanced system and identifies two particular content alignments to be studied: (1) items/tasks to models/targets and (2) test administered to students to blueprints. The first evaluates categorical concurrence and depth of knowledge, and the second evaluates balance of representation and range of knowledge. Outside the realm of alignment studies, but also planned, are a standards validation study and cognitive labs for technology- enhanced items.

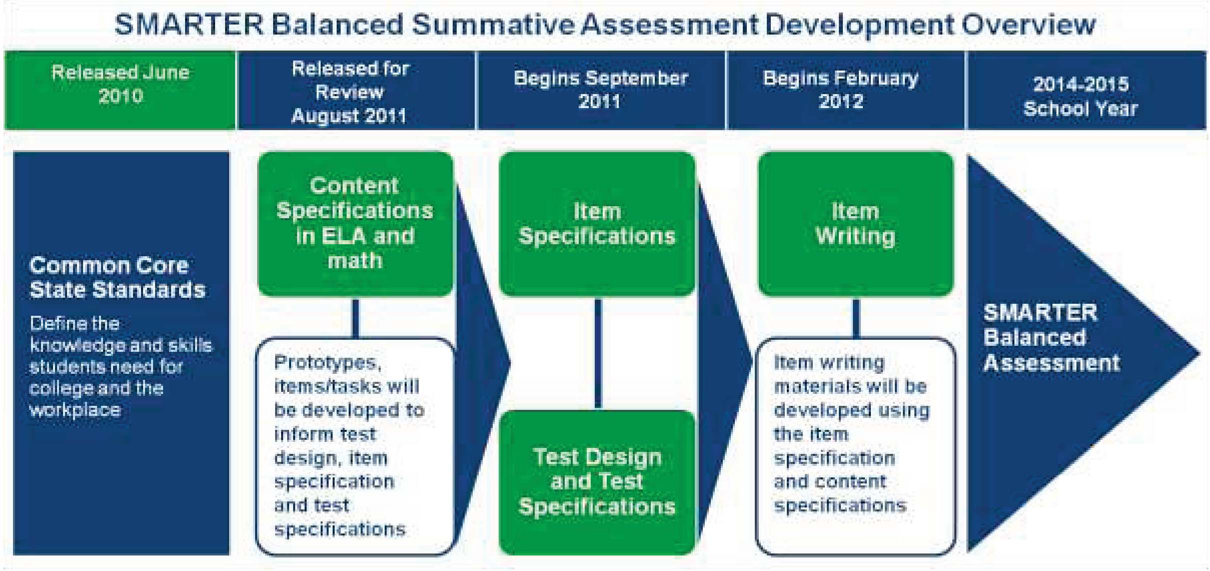
The Role of the Content Specifications

Developed in partnership with member states, leading researchers, content experts, and the authors of the Common Core, the Content Specifications are intended to ensure that the assessment system accurately assesses the full range the standards. The full Content Specifications for mathematics and ELA can be found online at this link:

http://www.smarterbalanced.org/smarter-balanced-assessments/

The Smarter Balanced *Content Specifications With Content Mapping for the Summative Assessment of the Common Core State Standards for Mathematics* and *Content Specifications With Content Mapping for the Summative Assessment of the Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects* provide clear and rigorous prioritized assessment targets that will be used to translate the grade-level Common Core standards into content frameworks along a learning continuum. From these content specifications the item/task specifications have been drafted. Assessment evidence at each grade level provides item and task specificity and clarifies the connections between instructional processes and assessment outcomes. The figure below, from the Smarter Balanced content specifications work shows the timeline for development and relationship between the CCSS, content specifications, item specifications, test design, item writing, and the Smarter Balanced Assessments.

Figure 2.


SMARTER Balanced Summative Assessment Development Overview
Released June 2010
Released for Review August 2011
Begins September 2011
Begins February 2012
2014-2015 School Year
Common Core State Standards
Define the knowledge and skills students need for college and the workplace
Content Specifications in ELA and math
Prototypes, items/tasks will be developed to inform test design, item specification and test specifications
Item Specifications
Test Design and Test Specifications
Item Writing
Item writing materials will be developed using the item specification and content specifications
SMARTER Balanced Assessment

Claims and Evidence for CCSS English Language Arts & Literacy Assessment

Defining Assessment Claims and Sufficient Evidence. The theory of action articulated by the Consortium illustrates the vision for an assessment system that will lead to inferences that ensure that all students are well-prepared for college and careers after high school. "Inference is reasoning from what one knows and what one observes, to explanations, conclusions, or predictions. One attempts to establish the weight and coverage of evidence in what is observed" (Mislevy, 1995, p 2).

"Claims" are the broad statements of the assessment system's learning outcomes, each of which requires evidence that articulates the types of data/observations that will support interpretations of competence towards achievement of the claims. A first purpose of this document is to identify the critical and relevant claims that will "identify the set of knowledge and skills that is important to measure for the task at hand" (NRC, 2001).

In close collaboration with content and technical experts, Consortium work groups and staff, and authors of the CCSS, the Content Specifications proposes claims for learning. E**ach claim is explained with a rationale describing the importance of the learning (embedded in the claim) in preparing students for college and careers.**

Relevant and sufficient evidence needs to be collected in order to support each claim. This can be accomplished using a variety of assessment items and tasks applied in different contexts. Data collection for the Smarter Balanced assessments is designed to be used to measure and make interpretations about within- and across-year student progress. The sufficient evidence section includes, for each claim, a brief analysis of the assessment issues to be addressed to ensure accessibility to the assessment for all students. **Each claim is accompanied with a description of the sufficient relevant evidence from which to draw inferences or conclusions about learning.**

Assessment Targets. For each Claim, a set of "Assessment Targets" are provided. Based on the description of sufficient evidence necessary to support each claim, the assessment targets describe the expectations of what will be assessed by the items and tasks within each claim. These summative assessment targets (evidence) at each grade level represent the prioritized content for summative assessment, and will be used to develop more detailed item and task descriptions through the item specification process.

The Purpose of the Item Specifications

The *Smarter Balanced Item and Task Specifications* are a bridge between the *Smarter Balanced Content Specifications* and the summative assessments through item/task development. The primary purpose of the *Smarter Balanced Item and Task Specifications* is to guide item/task developers in the work of designing items and tasks to illicit evidence from students tied to specific aspects of the content specifications. Another of the major purposes of the *Smarter Balanced Item and Task Specifications* is to help ensure that tests are measuring the intended constructs without contamination from construct-irrelevant factors. The Item and Task Specifications were written to help test developers distinguish between construct-relevant and irrelevant content, skills, and abilities.

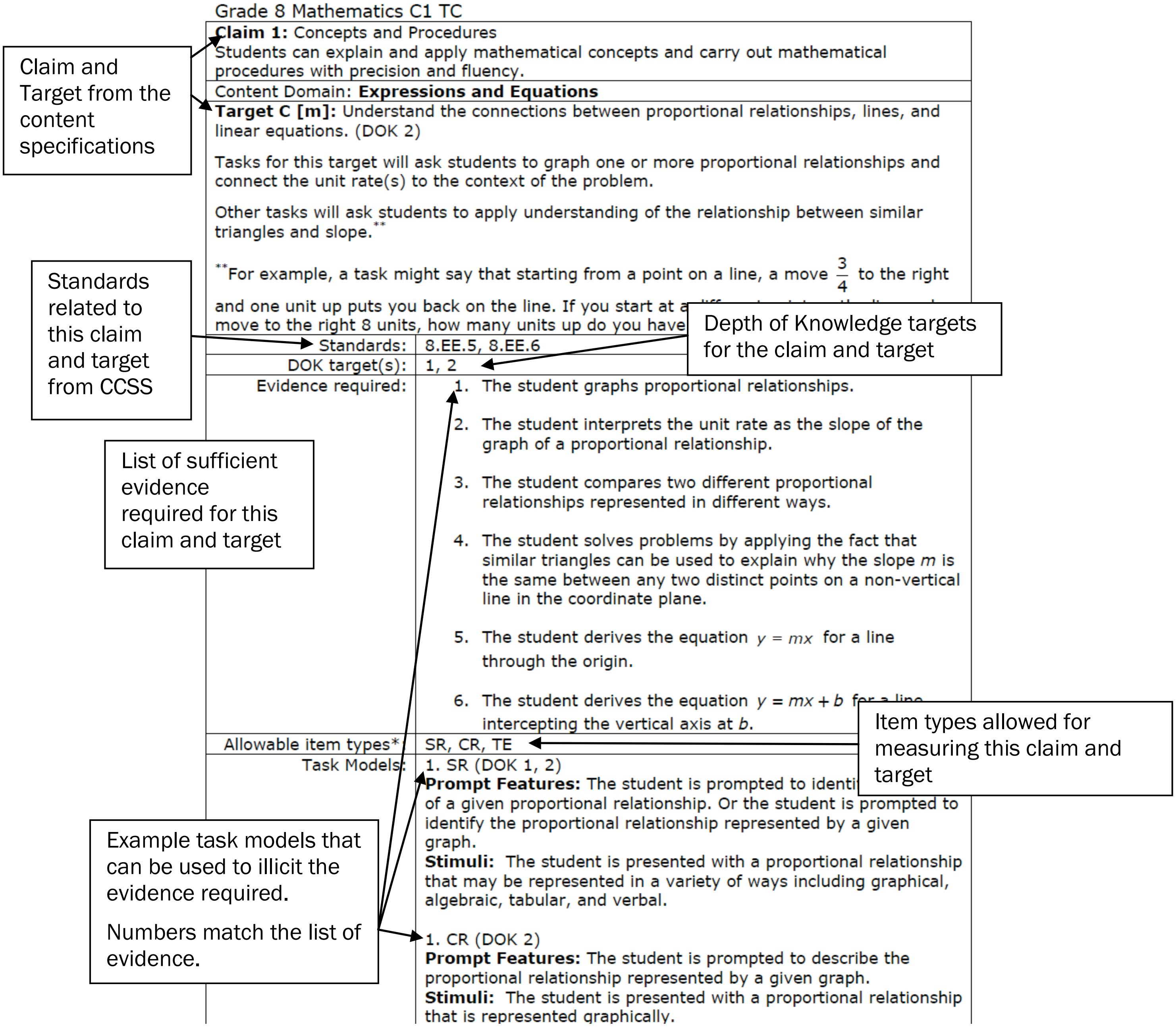
The major claims identified in the *Smarter Balanced Content Specifications for Mathematics and English Language Arts and Literacy* serve as the foundation for the *Item and Task Specifications*. The item specifications themselves emphasize the assessment targets articulated in the content specifications, and an item specification table is provided for each assessment target. Each item specification table will provide item writers with a clear definition of the construct intended to support the claim by identifying the sufficient evidence students will need to demonstrate to establish their skills and understanding. For each assessment target, the item specifications define the parameters/characteristics of items and tasks that will elicit evidence that is appropriate to support one or more claims about the assessment target…

The item specifications are expected to be iterative in nature and they will necessarily change over time, particularly as they relate to advances in technology and artificial intelligence scoring capability. As the assessment development process is refined, it is expected that the item specification tables will be edited to reflect current thinking and knowledge.

Sample items have been developed to accompany the item specification tables. The sample items illustrate how the task models can be operationalized in an assessment item.

The two figures below shows an example of an item specification table.

Figure 3. Item Specification Table Example.



Grade 8 Mathematics C1 TC

Claim 1: Concepts and Procedures
Students can explain and apply mathematical concepts and carry out mathematical procedures with precision and fluency.

Content Domain: **Expressions and Equations**

Target C [m]: Understand the connections between proportional relationships, lines, and linear equations. (DOK 2)

Tasks for this target will ask students to graph one or more proportional relationships and connect the unit rate(s) to the context of the problem.

Other tasks will ask students to apply understanding of the relationship between similar triangles and slope.**

**For example, a task might say that starting from a point on a line, a move $\frac{3}{4}$ to the right and one unit up puts you back on the line. If you start at a move to the right 8 units, how many units up do you have

Standards:	8.EE.5, 8.EE.6
DOK target(s):	1, 2
Evidence required:	1. The student graphs proportional relationships. 2. The student interprets the unit rate as the slope of the graph of a proportional relationship. 3. The student compares two different proportional relationships represented in different ways. 4. The student solves problems by applying the fact that similar triangles can be used to explain why the slope m is the same between any two distinct points on a non-vertical line in the coordinate plane. 5. The student derives the equation $y = mx$ for a line through the origin. 6. The student derives the equation $y = mx + b$ for a line intercepting the vertical axis at b.
Allowable item types*	SR, CR, TE
Task Models:	1. SR (DOK 1, 2) **Prompt Features:** The student is prompted to identi of a given proportional relationship. Or the student is prompted to identify the proportional relationship represented by a given graph. **Stimuli:** The student is presented with a proportional relationship that may be represented in a variety of ways including graphical, algebraic, tabular, and verbal. 1. CR (DOK 2) **Prompt Features:** The student is prompted to describe the proportional relationship represented by a given graph. **Stimuli:** The student is presented with a proportional relationship that is represented graphically.

Claim and Target from the content specifications

Standards related to this claim and target from CCSS

Depth of Knowledge targets for the claim and target

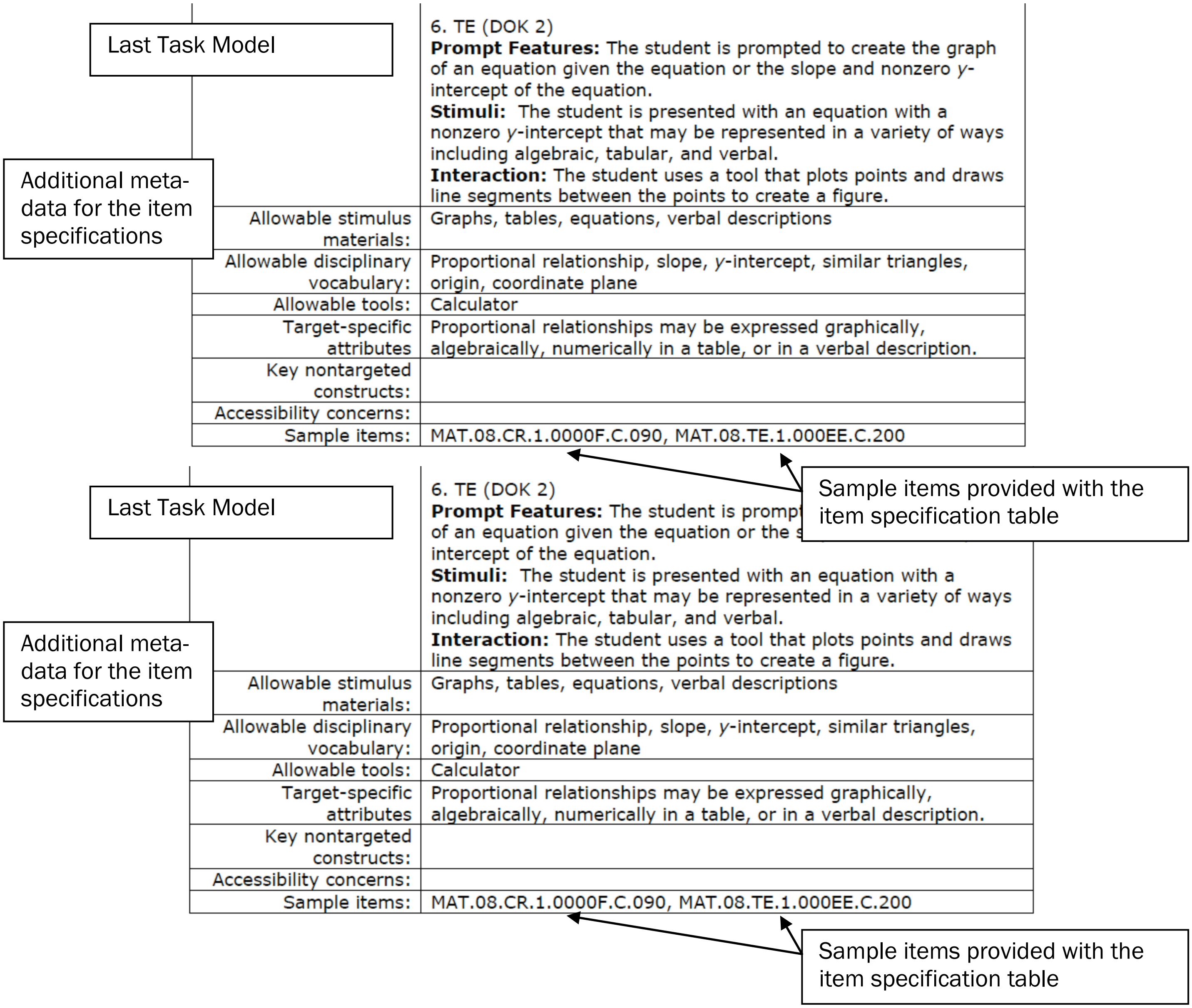
List of sufficient evidence required for this claim and target

Item types allowed for measuring this claim and target

Example task models that can be used to illicit the evidence required.

Numbers match the list of evidence.

Figure 4. Item Specification Table Example continued.



Last Task Model

Additional meta-data for the item specifications

	6. TE (DOK 2) **Prompt Features:** The student is prompted to create the graph of an equation given the equation or the slope and nonzero *y*-intercept of the equation. **Stimuli:** The student is presented with an equation with a nonzero *y*-intercept that may be represented in a variety of ways including algebraic, tabular, and verbal. **Interaction:** The student uses a tool that plots points and draws line segments between the points to create a figure.
Allowable stimulus materials:	Graphs, tables, equations, verbal descriptions
Allowable disciplinary vocabulary:	Proportional relationship, slope, *y*-intercept, similar triangles, origin, coordinate plane
Allowable tools:	Calculator
Target-specific attributes	Proportional relationships may be expressed graphically, algebraically, numerically in a table, or in a verbal description.
Key nontargeted constructs:	
Accessibility concerns:	
Sample items:	MAT.08.CR.1.0000F.C.090, MAT.08.TE.1.000EE.C.200

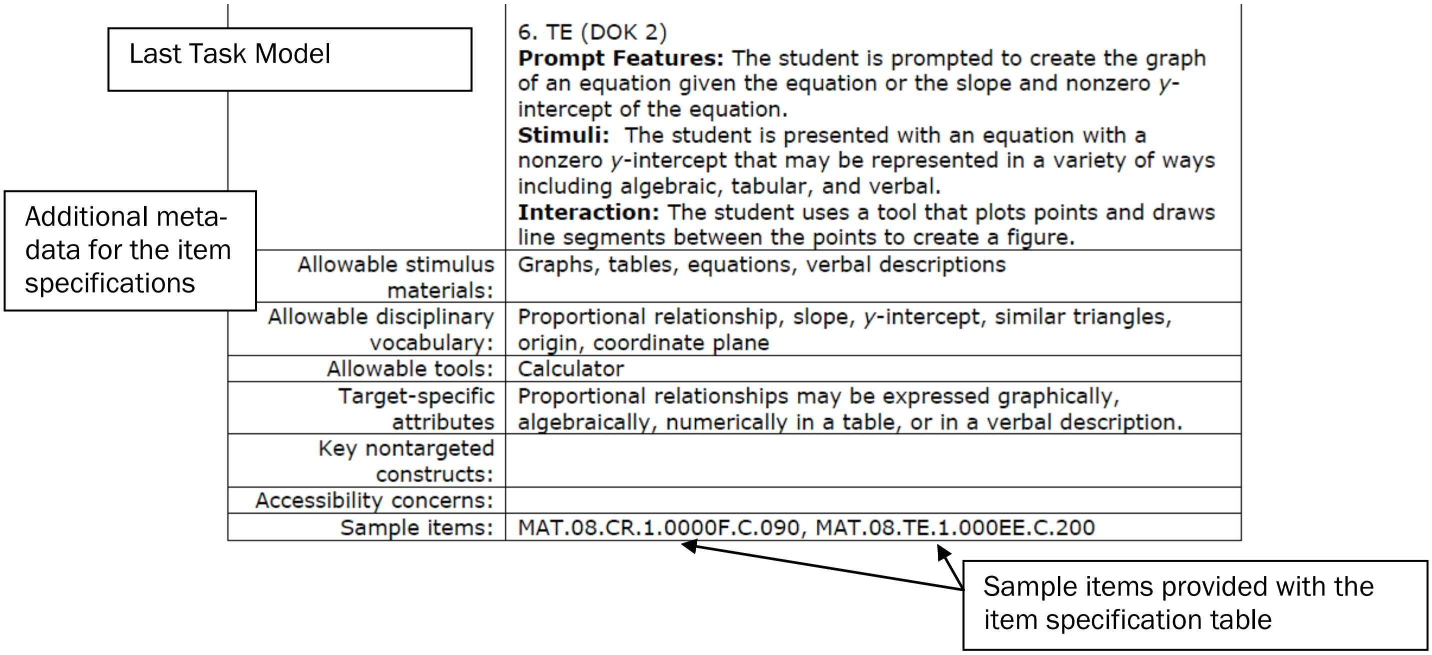
Sample items provided with the item specification table

Last Task Model

Additional meta-data for the item specifications

	6. TE (DOK 2) **Prompt Features:** The student is prompted to create the graph of an equation given the equation or the slope and nonzero *y*-intercept of the equation. **Stimuli:** The student is presented with an equation with a nonzero *y*-intercept that may be represented in a variety of ways including algebraic, tabular, and verbal. **Interaction:** The student uses a tool that plots points and draws line segments between the points to create a figure.
Allowable stimulus materials:	Graphs, tables, equations, verbal descriptions
Allowable disciplinary vocabulary:	Proportional relationship, slope, *y*-intercept, similar triangles, origin, coordinate plane
Allowable tools:	Calculator
Target-specific attributes	Proportional relationships may be expressed graphically, algebraically, numerically in a table, or in a verbal description.
Key nontargeted constructs:	
Accessibility concerns:	
Sample items:	MAT.08.CR.1.0000F.C.090, MAT.08.TE.1.000EE.C.200

Sample items provided with the item specification table



Last Task Model

Additional meta-data for the item specifications

	6. TE (DOK 2) **Prompt Features:** The student is prompted to create the graph of an equation given the equation or the slope and nonzero *y*-intercept of the equation. **Stimuli:** The student is presented with an equation with a nonzero *y*-intercept that may be represented in a variety of ways including algebraic, tabular, and verbal. **Interaction:** The student uses a tool that plots points and draws line segments between the points to create a figure.
Allowable stimulus materials:	Graphs, tables, equations, verbal descriptions
Allowable disciplinary vocabulary:	Proportional relationship, slope, *y*-intercept, similar triangles, origin, coordinate plane
Allowable tools:	Calculator
Target-specific attributes	Proportional relationships may be expressed graphically, algebraically, numerically in a table, or in a verbal description.
Key nontargeted constructs:	
Accessibility concerns:	
Sample items:	MAT.08.CR.1.0000F.C.090, MAT.08.TE.1.000EE.C.200

Sample items provided with the item specification table

The Role of the Companion Documents

Bias and Sensitivity Guidelines

The purpose of the *Smarter Balanced Assessment Consortium Bias and Sensitivity Guidelines* is to help ensure that the Smarter Balanced assessments are fair for all groups of test takers, despite differences in characteristics including, but not limited to, physical ability, ethnic group, gender, regional background, native language, race, religion, and socioeconomic status.

The goal of fairness in assessment can be approached by ensuring that test materials are as free as possible of unnecessary barriers to the success of diverse groups of test takers. The *Bias and Sensitivity Guidelines* document describes in detail how to avoid such barriers in the Smarter Balanced assessments. The *Bias and Sensitivity Guidelines* document is used in the design and development of the Smarter Balanced assessments, particularly in item writing and review.

The *Bias and Sensitivity Guidelines* document describes the rules agreed upon by the Smarter Balanced Assessment Consortium states for achieving fairness in test content. Only items that are in compliance with the *Guidelines* will be included in the Smarter Balanced assessments. Therefore, the *Guidelines* will help ensure that the test content is fair for test takers as well as acceptable to the many stakeholders and constituent groups within the Smarter Balanced states.

Accessibility and Accommodations Guidelines

The *Smarter Balanced Assessment Consortia Accessibility and Accommodation Guidelines* consist of six documents that are intended to be used by item writers and accessibility experts to make items and tasks accessible to as many students as possible. The guidelines combine best practices that have guided the development of paper-based tests for many years with recent advances facilitated by digital-delivery of assessment instruments. The methodology used to develop accessibility guidelines employed a collaborative development process that tapped expertise within the Smarter

Balanced Accessibility and Accommodation Work Group, Measured Progress, ETS, American Print House for the Blind, and an external sign language expert.

The table below summarizes the audience and intended use for each of the six accessibility guideline documents. The *General Accessibility Guidelines* and *English Language Learner Accessibility Guidelines* present research and best practices that are common to the application of Universal Design principals to item writing. The information presented in the *General Accessibility Guidelines* document is intended to be applied to the development of all items and tasks and addresses general issues that influence the accessibility of assessment content. The *English Language Learner Accessibility Guidelines* provide detailed information on accessibility issues specific to English language learners. When applied appropriately, these guidelines help assure assessment items and tasks are accessible for a broad spectrum of students.

As explained in the section on Universal Design in this document, the Access by Design model articulates, meeting the access needs for some students requires that item and task content be presented using a specific representational form in order to adequately stimulate the construct of interest. The *Tactile/Braille Accessibility Guidelines*, *Mathematics Audio Guidelines*, *ELA Audio Guidelines*, and *Sign Language Guidelines* provide information that accessibility experts can use to create item extensions for different representational forms. Each type of accessibility information is designed to provide information that supplements the item content, so that an accessibility need does not interfere with the measure of the intended construct. These guidelines help ensure that accessibility information is specified consistently to provide high-quality access.

Table 3: Accessibility Guidelines Documents

Guideline Document	Audience	Intended Use
General Accessibility Guidelines	Item and Task Writers	Create items and tasks that are accessible to as many students as possible.
English Language Learner Accessibility Guidelines	Item and Task Writers	Create items and tasks that are accessible to as many students as possible.
Tactile/Braille Accessibility Guidelines	Tactile/Braille Experts	Create tactile and braille item extensions to items and tasks.
Mathematics Audio Guidelines	Audio Experts	Create mathematics audio extensions to items and tasks.
English Language Arts Audio Guidelines	Audio Experts	Create ELA audio extensions to items and tasks.
Sign Language Guideline	Sign Language Experts	Create sign language extensions to items and tasks.

Style Guide

The *Smarter Balanced Style Guide* provides style conventions and specifications for both test items and test forms created for the Smarter Balanced Assessment Consortium. Addressing a wide range of topics, the style guide contains global conventions for test items and test forms, content-specific conventions for English language arts and mathematics, specific conventions for graphics and technology-enhanced items, guidelines for grammar and usage, and specifications for printed test forms. The style guide also explains the reasoning behind some of the more significant conventions.

The style guide serves as a resource for content specialists, item writers, editors, graphic designers, and other individuals involved in developing and producing content for Smarter Balanced assessments. The style guide is a comprehensive document; however, because it is not possible to anticipate all issues that may arise during item development, the style guide provides a list of recommended resources that can also be consulted.

Stimulus Specifications

The *Smarter Balanced Stimulus Specifications* document provides detail that will help item writers select appropriate topics, features, and layouts for stimulus material selected or written for ELA/Literacy item design. The parameters presented are informed by best practices described in the Common Core State Standards (CCSS), the *Smarter Balanced Assessment Consortium Content Specifications for ELA*, and the practices shown in Smarter Balanced states' guidelines. Appropriate kinds of texts, grade level- appropriate topics and complexity, and other features pertinent to the domain of ELA is examined and guidance provided. Included in these specifications is a section on measures to determine text complexity that provides guidance for using quantitative and qualitative measures to evaluate grade-level texts for inclusion in Smarter Balanced assessments. Guidance on stimulus for mathematics items can be found in the mathematics material.

General Considerations in Item Development

The next several sections of the document address general considerations in item development, including sections on Cognitive Complexity, Universal Design, Grade Appropriateness, Vocabulary and Language, and Artificial Intelligence Scoring

Cognitive Complexity

In addition to attending to the sufficient evidence required to establish student achievement as specified by the Smarter Balanced assessment claims and targets, item writers additionally must consider the breadth and depth of knowledge communicated by these targeted expectations. An aligned assessment must include items and tasks requiring the highest level of cognitive complexity prescribed by the assessment claims and targets. The Smarter Balanced Assessment Consortium has adopted a Cognitive Rigor Matrix for its assessment program. This matrix draws from two widely accepted measures to describe cognitive rigor: Bloom's (revised) Taxonomy of Educational Objectives and Webb's Depth-of-Knowledge Levels. The Cognitive Rigor Matrix has been developed to integrate these two models as a strategy for analyzing instruction, for influencing teacher lesson planning, and for designing assessment items and tasks. (To download the full article describing the development and uses of the Cognitive Rigor Matrix and other support materials, go to: http://www.nciea.org/publications/cognitiverigorpaper_KH11.pdf).

The Common Core State Standards require high-level cognitive demand, such as requiring students to demonstrate deeper conceptual understanding through the application of content knowledge and skills to new situations and sustained tasks. For each Assessment Target in English language arts and mathematics, the depth(s) of knowledge (DOK) that the student needs to bring to the item/task has been identified. Short descriptions of the Cognitive Rigor Matrix for each content area are provided below.

Table 4. A "Snapshot" of the Cognitive Rigor Matrix for Mathematics.

Depth of Thinking (Webb) + Type of Thinking (Revised Bloom)	DOK Level 1 Recall & Reproduction	DOK Level 2 Basic Skills & Concepts	DOK Level 3 Strategic Thinking & Reasoning	DOK Level 4 Extended Thinking
Remember	• Recall conversions, terms, facts			
Understand	• Evaluate an expression • Locate points on a grid or number on number line • Solve a one-step problem • Represent math relationships in words, pictures, or symbols	• Specify, explain relationships • Make basic inferences or logical predictions from data/observations • Use models /diagrams to explain concepts • Make and explain estimates	• Use concepts to solve non-routine problems • Use supporting evidence to justify conjectures, generalize, or connect ideas • Explain reasoning when more than one response is possible • Explain phenomena in terms of concepts	• Relate mathematical concepts to other content areas, other domains • Develop generalizations of the results obtained and the strategies used and apply them to new problem situations
Apply	• Follow simple procedures • Calculate, measure, apply a rule (e.g.,rounding) • Apply algorithm or formula • Solve linear equations • Make conversions	• Select a procedure and perform it • Solve routine problem applying multiple concepts or decision points • Retrieve information to solve a problem • Translate between representations	• Design investigation for a specific purpose or research question • Use reasoning, planning, and supporting evidence • Translate between problem & symbolic notation when not a direct translation	• Initiate, design, and conduct a project that specifies a problem, identifies solution paths, solves the problem, and reports results
Analyze	• Retrieve information from a table or graph to answer a question • Identify a pattern/trend	• Categorize data, figures • Organize, order data • Select appropriate graph and organize & display data • Interpret data from a simple graph • Extend a pattern	• Compare information within or across data sets or texts • Analyze and draw conclusions from data, citing evidence • Generalize a pattern • Interpret data from complex graph	• Analyze multiple sources of evidence or data sets

Evaluate			• Cite evidence and develop a logical argument • Compare/ contrast solution methods • Verify reasonableness	• Apply understanding in a novel way, provide argument or justification for the new application
Create	• Brainstorm ideas, concepts, problems, or perspectives related to a topic or concept	• Generate conjectures or hypotheses based on observations or prior knowledge and experience	• Develop an alternative solution • Synthesize information within one data set	• Synthesize information across multiple sources or data sets • Design a model to inform and solve a practical or abstract situation

(Hess, Carlock, Jones, & Walkup, 2009)

Table 5. A "Snapshot" of the Cognitive Rigor Matrix for English Language Arts.

Depth of Thinking (Webb) + Type of Thinking (Revised Bloom)	DOK Level 1 Recall & Reproduction	DOK Level 2 Basic Skills & Concepts	DOK Level 3 Strategic Thinking & Reasoning	DOK Level 4 Extended Thinking
Remember	• Recall, locate basic facts, definitions, details, events			
Understand	• Select appropriate words for use when intended meaning is clearly evident	• Specify, explain relationships • Summarize • Identify central ideas	• Explain, generalize, or connect ideas using supporting evidence (quote, text evidence, example…)	• Explain how concepts or ideas specifically relate to other content domains or concepts
Apply	• Use language structure (pre/suffix) or word relationships (synonym/antonym) to determine meaning	– Use context to identify word meanings • - Obtain and interpret information using text features	• Use concepts to solve non-routine problems	• Devise an approach among many alternatives to research a novel problem
Analyze	• Identify the kind of information contained in a graphic, table, visual, etc.	– Compare literary elements, facts, terms, events • – Analyze format, organization, & text structures	• Analyze or interpret author's craft (e.g., literary devices, viewpoint, or potential bias) to critique a text	• Analyze multiple sources or texts • Analyze complex/ abstract themes
Evaluate			• Cite evidence and develop a logical argument for conjectures based on one text or problem	• Evaluate relevancy, accuracy, & completeness of information across texts/ sources

Create	- - Brainstorm ideas, concepts, problems, or perspectives related to a topic or concept	- -Generate conjectures or hypotheses based on observations or prior knowledge and experience	- Develop a complex model for a given situation - Develop an alternative solution	- Synthesize information across multiple sources or texts - -Articulate a new voice, alternate theme, new knowledge or perspective

(Hess, Carlock, Jones, & Walkup, 2009)

Universal Design

The concept of Universal Design focuses on “the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design” (CUD, 1997). When applied to the development of assessment items and tasks, the concept of Universal Design aims to create items and tasks that accurately measure the targeted knowledge, skills, and abilities for all students. However, the concept of Universal Design recognizes that a single solution rarely, if ever, functions well for all users. For this reason, Universal Design also embraces the concept of allowing users to select from multiple alternatives. As Rose and Meyer emphasize, “Universal Design does not imply ‘one size fits all’ but rather acknowledges the need for alternatives to suit many different people’s needs…the essence of Universal Design is flexibility and the inclusion of alternatives to adapt to the myriad variations in learner needs, styles, and preferences” (Rose & Meyer, p. 4).

When developing assessment items and tasks, the spirit of Universal Design is captured by first applying the general guidelines to design items and tasks that work well for a broad range of students and then applying the accompanying guidelines to develop adaptations that extend the ability of an item or task to also accurately measure students with specialized access needs.

When applied to assessment items and tasks, Universal Design has two important implications. First, Universal Design requires item writers to consider the full range of students who are expected to be measured by an item or task and to design the item to function appropriately for the widest range of these students without adaptation. The item specifications and guidelines provide several considerations that can expand the range of students for which an item or task functions well. As an example, using vocabulary that is commonly used in school rather than vocabulary that is associated with specialized activities that may not be familiar to all students (e.g., sport-specific terminology such as “ski binding” or “putter,” hobby-specific vocabulary such as “yarn over” or “rabbet joint”) can improve the accuracy with which an item or task is able to stimulate the targeted knowledge, skill, and ability of students who are unfamiliar with such specialized vocabulary. Similarly, minimizing the use of visual materials such as figures, graphs, and maps to those cases when they are absolutely required by an item can improve an item or task’s functioning for students with visual needs and for students who have challenges processing multiple pieces of information.

Second, Universal Design requires item writers to create items that support adaptations that are designed to meet the needs of specific subgroups of students. As an example, minimizing the complexity of visual materials so that they can be described verbally or represented as a tactile image supports the adaptations of that content for students with visual needs.

Valid assessment of student knowledge, skills, and abilities requires a two-way communication between an assessment item and a student that involves three critical steps. The first step in this communication process focuses on presenting information to a student in order to activate or stimulate the knowledge, skill, or ability that is the target of assessment. Second, the student is provided an opportunity to interact with content that is presented by an item as s/he applies the targeted knowledge, skill, or ability. Third, the student provides evidence about their knowledge, skill, or ability through their response to the assessment item or task. It is through this three-step process that an assessment item or task attempts to access the targeted knowledge, skills, or abilities that operate within the student.

Access by Design is an approach to developing items and tasks that aims to improve the accuracy with which assessment items and tasks measure targeted knowledge, skills, and abilities by maximizing the range of students for which an item accurately stimulates the assessment target, allows the student to interact with content as they apply their knowledge, skills, and abilities, and enables students to produce responses that accurately reflect the outcome of their thinking. Maximizing the range of students for which items and tasks provide valid measures of the target of assessment involves a three-step process.

The first step, which is a core component of Evidence Centered Design, is to clearly define the knowledge, skills, and/or abilities that are the target of assessment. The Smarter Balanced Assessment Consortium uses the term “assessment target” to refer to the knowledge, skills, and/or abilities that are the target of assessment. When defining an assessment target, it is critical to clearly articulate the knowledge, skill, or ability that is intended to be measured. As part of this process, it is important to consider what knowledge, skill, and ability the student must bring to the item in order to succeed, and what knowledge, skills, or abilities are not intended to be measured. As an example, a mathematics item that asks student to perform addition with two digits in the context of a real-world problem might require the student to bring to the item knowledge of addition, knowledge of the number system, and an ability to relate real-world situations to appropriate mathematical operations. However, this item might not intend to measure a student’s ability to read print-based text. Clearly defining assessment targets and carefully considering what is and is not intended to be measured is an essential first step in maximizing the validity of assessment.

The second step focuses on applying principles of Universal Design to the design and authoring of the content that forms each assessment item and task. The third step involves providing extensions to assessment content in order to better meet specific accessibility needs. One example of an extension is specifying how text-based content is to be presented in braille form. Key to providing extensions, however, is careful consideration of whether accessibility supports provided through an extension infringe on the knowledge, skills, and/or ability that is the target of assessment. When this occurs, students may be better able to access the item, but the item no longer provides a valid measure of the assessment target. Together, the application of principles of Universal Design and the use of extensions designed to meet specific access needs are the foundation of Access by Design.

While the goal of applying principles of Universal Design is to develop items that function well for all students, the Access by Design model recognizes that extensions to item content may be necessary to maximize the range of students for which an item or task accurately measures the targeted knowledge, skills, and abilities.

Grade Appropriateness

The Item and Task Specifications for both Mathematics and English Language Arts and Literacy provide guidance on determining the grade appropriateness of stimulus materials and the contexts used for items. The CCSS specifies rigorous expectations for a student’s ability to read and comprehend complex material, and the item specifications developed for the Smarter Balanced summative assessment have followed this lead. Please refer to the Stimulus Specifications, for more detail on determining complexity level of stimulus material specifically for ELA/Literacy items. For more information on CCSS expectations for text complexity, see Appendix A of the CCSS for English Language Arts and Literacy.

Whenever an assessment item is measuring grade-level specific content, vocabulary and skills identified for that grade level are appropriate to include. For development of the context for assessment items and tasks, item writers are encouraged to use rich and complex texts and contexts that would be expected through the previous grade level to allow for access to the greatest number of students.

Vocabulary and Language

The Common Core State Standards (CCSS) articulate through the College and Career Readiness Anchor Standards for Language (National Governors' Association Center for Best Practices & Council of Chief State School Officers, 2010, p. 25) an expectation that students will build a foundation for college and career readiness in language. The CCSS further convey expectations for students' mastery of standard English grammar, usage, and mechanics and an expectation of their utilization of these language skills as well as developing effective skills in using language to convey meaning. These skills are built as students progress through the grades, with an emphasis on integration of these skills into other content areas. As such, students' vocabulary and sophistication of language use is expected to increase over time. To do this, students must use their acquired language skills in the content areas they are studying.

Another key expectation of the CCSS is the ability to determine or clarify the meaning of grade-appropriate words encountered through study in the content areas including an understanding that words sometimes have nonliteral meanings, shades of meaning, and different meanings depending on context of use.

The Smarter Balanced Item and Task Specifications facilitate these CCSS expectations by encouraging use of content-specific language appropriate to the grade as articulated within the CCSS. Students are expected to have mastered vocabulary and language from previous grade levels, as well as use the vocabulary and language of the current grade level when being assessed on content specific to their current grade level. To avoid construct irrelevance from language, when being assessed on non-content-specific material, it is appropriate to use vocabulary or language from previous grade levels. For example, when assessing Writing in grade 8, language and reading skills through grade 7 are expected.

For English language learner students taking a large-scale content assessment, there is concern that the use of language and vocabulary on the assessment can limit accessibility for the student, thus compromising the validity of the test score interpretation for that student. On the other hand, if language and vocabulary use are limited to accommodate such learners, the validity of the assessment with respect to the content domain can be compromised. For these reasons, it is important that test developers pay attention to the main threat to validity when assessing content knowledge, which stems from language factors that are not relevant to the construct being assessed. To assist item writers and test developers with this challenging effort, *Guidelines for Accessibility for English Language Learners* has been included as a companion document to the *Item and Task Specifications.*

Artificial Intelligence Scoring

Artificial intelligence (AI) scoring is a type of automated scoring procedure in which computer algorithms are used to score response types that are open-ended enough that they cannot be scored by means of simple rules or other deterministic procedures. In particular, it excludes typical technology-enhanced response types such as drag-and-drop items, hotspots, or texts highlighting though these also require substantial engineering in order to function correctly.

The successful application of AI scoring for an assessment is doubly dependent on the types of tasks designed.

First, it depends on the tasks being created in such a manner that the evidence of students' knowledge, skills, and attributes of interest can be reliably identified by automated means from student response characteristics. What this entails will differ somewhat from task type to task type, but it generally means that the class of responses to be given credit for an item will need to be both constrained, and defined in terms of characteristics that can be operationalized by the computer. The more response variability is licensed, and the less amenable to computer modeling the item rubric is, the more difficult it will be to apply AI scoring.

Second, how automated scoring can be applied to an item type depends on how the interface specifications for that item type define the way in which students will interact with stimulus material and encode their responses for processing. In particular, where a task has multiple parts, the response to each part should be provided in a separate entry field, insofar as possible. Furthermore, the tools for authoring and encoding special types of responses (such as equations) should be designed so that they are easy for students to use, and constrained in order to prevent ambiguous or uninterpretable responses.

The item types under current consideration within these Item Specifications occupy various points along the continuum of difficulty for AI scoring methods, consistent with the Consortium's desire to ensure broad and deep representation of all targeted proficiencies and standards, and to drive innovation beyond the current state of the art in AI scoring technology. In the mathematics content area, items designed to elicit mathematical graphs, equation, or expressions are generally suitable for automated scoring in its current state, although test delivery considerations must be attended to carefully. In English language arts, the extended writing tasks are in most respects suitable for existing AI scoring capabilities, but will likely need to be augmented by human judgment where it is necessary to assess whether a text is well reasoned or well suited to the intended audience. For both content areas, the short, textual constructed-response items represent the greatest challenge for state-of-the-art capabilities for AI scoring at the present time. Because they are intended to be scored based on general evidence of students' reasoning processes rather than referencing of specific concepts, accurate and valid AI scoring for these items will require significant technological advances.

In order to mitigate the risk associated with incorporating tasks that are challenging for AI scoring into the assessment design, the Smarter Balanced Consortium may consider constraining selected item types to better accommodate the technology.

General Considerations According to Item Type

The Smarter Balanced Item and Task Specifications Documents provide detailed requirements for writing five types of items and tasks: selected response items, constructed response items, extended response items, technology-enhanced items, and performance tasks. The sections below outline the requirements for each item type, with additional detail proved in the content-specific item specifications.

Selected Response Items

Selected Response Items (SR) contain a series of options from which to choose correct responses. The Consortium's emphasis will be on the development of items that reflect important knowledge and skills consistent with the expectations of the CCSS across the Depths of Knowledge (i.e., Recall/Literal Comprehension, Interpretation/Application, and Analysis/Evaluation). Carefully constructed and reviewed selected response items will allow students to demonstrate their use of complex thinking skills, such as formulating comparisons or contrasts; identifying cause and effects; identifying patterns or conflicting points of view; categorizing, summarizing, or interpreting information. The appropriate and judicious use of selected response items provides for a cost-effective means to address content in terms of test development, administration, and scoring.

Selected Response (SR) items will measure one or more content standard(s). A single SR item will not measure content standards in both mathematics and English language arts. For selected response items that are multiple choice, there will be up to four possible answer options (e.g., one correct answer and three wrong answer choices [distractors]). Selected response items should include, but not be limited to, multiple-choice items.

The following list of considerations for item writing should be addressed when developing selected response items.

- Each selected response item should be written to focus primarily on one assessment target. Secondary targets are acceptable and are listed in the item meta-data of sample items as appropriate, but it should be clear to all stakeholders which assessment target is the focal point of the item.
- Items should be appropriate for students in terms of grade-level difficulty, cognitive complexity, and reading level. For non-reading items, the reading level should be approximately one grade level below the grade level of the test, except for specifically assessed terms or concepts.
- Items are expected to include concepts detailed in the CCSS of lower grades.
- Items should provide clear and complete instructions to students.
- Each item should be written to clearly elicit the desired evidence of a student's knowledge, skills, or abilities.
- Options should be arranged according to a logical order whenever possible (e.g., alphabetical, least to greatest value, greatest to least value, length of options).

Constructed Response and Extended Response Items

Constructed Response (CR) is a general term for items requiring the student to generate a response as opposed to selecting a response. Both short and extended constructed response items will be used. Short constructed response items may require test-takers to enter a single word, phrase, sentence, number, or set of numbers, whereas extended constructed response items will require more elaborated answers and explanations of reasoning. These kinds of constructed response items

> … allow students to demonstrate their use of complex thinking skills such as formulating comparisons or contrasts; proposing cause and effects; identifying patterns or conflicting points of view; categorizing, summarizing, or interpreting information; and developing generalizations, explanations, justifications, or evidence-based conclusions (Darling-Hammond & Pecheone, 2010). These complex thinking skills are consistent with the expectations for college and career readiness and will be included in both the English language arts and mathematics assessments. (Smarter Balanced Assessment Consortium RTTT Proposal, p. 53.)

Constructed response items will measure one or more content standard(s). A single CR item will not measure content standards in both mathematics and English language arts. It is expected that constructed response items will generally be scored by computer, using Artificial Intelligence (AI) models as appropriate, with human backup scoring for validation.

In mathematics, a specific constructed response item type designated as extended response (ER) will be employed. ER items/tasks will contribute to the performance task component; CR items will contribute to the computer-adaptive component. Therefore, in mathematics a CR is a brief constructed-response item that focuses on a particular skill or concept and will be included in the computer-adaptive component. The length of time these CRs take to administer should typically vary from 1 to 5 minutes. An ER item/task is designed to cover content at a greater depth than a regular CR item. The time allotted to administer ER items/tasks should vary from 5 to 20 minutes.

The following list of considerations for item writing should be addressed when developing constructed response items.

- Each item/task should be written to assess a primary content domain as identified in the assessment targets of the specified grade. Secondary content domains are also possible and should be listed in order of prominence when completing the item form.
- Items/tasks should be appropriate for students in terms of grade-level difficulty, cognitive complexity, and reading level. For non-reading items, the reading level should be approximately one grade level below the grade level of the test, except for specifically assessed content terms or concepts.
- Items/tasks are expected to include concepts detailed in the CCSS of lower grades.
- At grades 6-8, all mathematics items/tasks should be written so they can be answered without using a calculator. However, some targets may require the use of an online calculator tool in order to efficiently problem solve. In these cases, the calculator tool will appear in the specification table under “allowable manipulative materials.”
- Items/tasks should provide clear and complete instructions to students.
- CR items/tasks should be written to clearly elicit the desired evidence of a student’s KSA.
- For CR items, a complete key and/or scoring rubric must be included with the item along with a justification for the solution, as needed.
- For ER items/tasks in mathematics, a Sample Top-Score Response must be provided, accompanied by a scoring rubric that details the rationale for awarding each score point in terms of the evidence demonstrated by a student’s response. Scoring guidelines for CRs, ERs, and PTs, are discussed more thoroughly in the content-specific item specifications documents.

Technology- Enhanced Items

Technology-Enhanced (TE) Items have been defined by the Smarter Balanced item Development Work Group as follows. Technology-Enhanced Items employ technology to:

- Elicit a response from the student (e.g., selecting one or more points on a graphic, dragging and dropping a graphic from one location to another, manipulating a graph)…, and/or
- TE Items employ technology to assess content, cognitive complexity, and Depth of Knowledge not assessable otherwise. Because of the cost in development, scoring, and ongoing calibration, Smarter Balanced will employ TE in situations in which static SR and static CR are inadequate.
- The ultimate goal of TE items is to provide better measurement of student knowledge and skills through technology.

The effective use of technology will expand the nature of the content that can be presented as well as the knowledge, skills, and processes that can be assessed (Quellmalz & Moody, 2004). Technology-Enhanced items will take advantage of drag-and-drop, hot spot, drawing, graphing, gridded-response items (which generally have numerical answers where students can key-in responses), and simulation technologies, along with the use of online tools to measure content that was previously not assessed or was assessed through constructed response item formats requiring more elaborate scoring procedures.

With the advent of online assessments, many capabilities now exist for multi-media stimuli, interactive reference materials, and richer, more interactive responses from students. Smarter Balanced is committed to using advanced technology capabilities when doing so allows a better measurement of what students know and can do. However, new technology should not be used for the sake of including technology not previously used. It should be used when previously established measurement tools prove inadequate to properly measure students' abilities.

Technology-Enhanced items will measure one or more content standard(s). A single TE item will not measure content standards in both mathematics and English language arts. However, the same kinds of technology-enhanced interactions could be used across content areas.

Technology-Enhanced (TE) items/tasks are desirable when they can provide evidence that could not be as reliably obtained from SR and CR items. Additionally, components of certain extended-response (ER) items (in mathematics) and performance tasks may employ TE tools as part of the task. An expressed desire on the part of the Consortium is that the use of TE items in the assessments will ultimately encourage classroom use of authentic mathematical computing tools (e.g., spreadsheets, interactive geometry software) as part of classroom instruction.

At this time items requiring Artificial Intelligence (advanced programming logic) are not included for consideration directly as part of a template specification, although they may be considered through the adoption of scoring engines as described in the Smarter Balanced statement on principles for adoption of Artificial Intelligence. Artificial Intelligence items also include items that require a dependence of one step of a process with another process (example: a student creates a specified closed object, then is asked to provide the area of that object, and credit should be given for the student getting the area correct even if they created the object incorrectly).

TEI Templates

Technology-e Enhanced item specifications make use of Template Specifications. A template describes a single interaction, the response data collected as a result of that interaction, and the logic applied to score the response data. Templates can support dichotomous or polytomous scoring. The intended audience for TEI Templates is software developers, item development managers, and assessment planners.

Templates are designed to be used to create items which use the specific parameters allowed by the template. By predefining the interactions and parameters of items, software development can be cost effective and efficient. TE templates allow for a greater amount of “complex” items to be scored automatically, so are more suitable in an adaptive assessment environment.

Response interactions described in TE templates can be used in conjunction with Selected Response and Constructed Response interactions. A multi-step task may involve using a combination of interactive media stimuli (referred to as Technology Enabled stimuli), Technology Enhanced responses, and selected/constructed responses. While technology-e Enhanced interactions could be used in isolation, it is not the expectation that they are only used in isolation.

Technology Enhanced templates focus on the response interactions beyond Selected Response and Construction response interactions. Items may contain Technology Enabled stimuli (multi-media stimuli like videos or audio recordings, or interactive media, like layered maps, simulations, or creative applets), which can be used with or without a Technology-Enhanced response interaction. However, TE templates do not describe or specify the construction or use of Technology Enabled stimuli.

Included in the item and task specifications are the templates currently under consideration, from which a small subset will be selected for the initial operational stages of Smarter Balanced TE items. These include templates thought to be useful to create items that will allow students to provide evidence of targets across Mathematics and English Language Arts, though they could also be used for other content areas. Smarter Balanced acknowledges the important intent to adopt innovative technology designs along with measurement procedures for effective validation and use of such TE items. Thus Smarter Balanced will focus on adopting technology as it becomes sufficiently mature for the intended use. Within this principle, new templates will be considered when new evidence is required for students to produce.

The templates are constructed to indicate a range of innovative response actions that may be made available in the assessments. Additional innovations in the prompt or directive through new media, such as including potentially brief full-motion animation if applicable or textual use only of standardized agents, may be introduced when consistent with the Smarter Balanced framing ideas above, but are not fully exemplified in the templates as they do not require new response actions indicated as the role of the templates.

The templates are constructed to indicate a range of innovative response actions that may be made available in the assessments. Additional innovations in the prompt or directive through new media, such as including potentially brief full-motion animation if applicable or textual use only of standardized agents, may be introduced when consistent with the Smarter Balanced framing ideas above, but are not fully exemplified in the templates as they do not require new response actions indicated as the role of the templates.

Technology-Enhanced (TE) items/tasks are desirable when they can provide evidence for that what could not be as reliably obtained from SR and CR items. Additionally, components of certain Extended-Response (ER) items (in mathematics) and performance tasks may employ TE tools as part of the task. An expressed desire on the part of the consortium is that the use of TE items in the assessments will ultimately encourage classroom use of authentic mathematical computing tools (e.g., spreadsheets, interactive geometry software) as part of classroom instruction.

Performance Tasks

Smarter Balanced defined Performance Tasks in their Race to the Top application as follows:

> [Performance tasks]…will provide a measure of the student's ability to integrate knowledge and skills across multiple [content] standards — a key component of college- and career readiness. Performance [tasks] will be used to better measure capacities such as depth of understanding, research skills, and complex analysis, which cannot be adequately assessed with [selected response] or constructed response items. (p. 42).

The Smarter Balanced Performance Task Work Group has identified the essential characteristics by specifying a performance task must:

- Integrate knowledge and skills across multiple content standards or English language arts strands/mathematics domains;
- Measure capacities such as depth of understanding, research skills, and/or complex analysis with relevant evidence;
- Require student-initiated planning, management of information and ideas, and/or interaction with other materials;
- Require production of more extended responses (e.g., oral presentations, exhibitions, product development), in addition to more extended written responses that might be revised and edited;
- Reflect a real-world task and/or scenario-based problem;
- Lend itself to multiple approaches;
- Represent content that is relevant and meaningful to students;
- Allow for demonstration of important knowledge and skills, including those that address 21st century skills such as critically analyzing and, synthesizing media texts;
- Focus on big ideas over facts;
- Allow for multiple points of view and interpretations;
- Require scoring that focuses on the essence of the task;
- Reflect one or more of the Standards for Mathematical Practice, Reading and Writing (or Speaking and Listening) processes; and
- Seem feasible for the school/classroom environment.

The general specifications for Smarter Balanced performance tasks build upon the work of the Smarter Balanced Performance Task Work Group, which provided guidelines describing the general characteristics desired in the performance tasks. The specifications address fifteen different aspects of performance tasks, such as overall structure, allowable teacher-student interactions, tools and other resources, and scoring requirements.

The task specifications call for multi-part, multi-session activities during which students individually will produce several scorable responses, products, or presentations. All this will be accomplished within controlled classroom settings. More detailed information regarding time requirements, stimulus materials, products, etc. is provided in subject-specific performance task specifications and in the target-specific tables describing performance task models. Specifications developed by groups other than the performance task team (specifications/guidelines for stimulus materials, rubrics, formatting and style, bias and sensitivity, and accessibility and accommodations) are not repeated in the general performance task specifications even though they apply.

As with TE items/tasks, an entire section of these *Specifications* contains information related to the development of high-quality performance tasks, and a writer must refer to that section when attempting to write these tasks. In short, performance tasks should:

- Integrate knowledge and skills across multiple claims and targets;
- Measure capacities such as depth of understanding, research skills, and/or complex analysis with relevant evidence;
- Require student-initiated planning, management of information/data and ideas, and/or interaction with other materials;
- Reflect a real-world task and/or scenario-based problem;
- Allow for multiple approaches;
- Represent content that is relevant and meaningful to students;
- Allow for demonstration of important knowledge and skills, including those that address 21st century skills such as critically analyzing and synthesizing information presented in a variety of formats, media, etc.;
- Require scoring that focuses on the essence of the Claim(s) and Targets for which the task was written. Scoring rules are described in detail in the Performance Task section of the content-specific item specifications documentation;
- Be feasible for the school/classroom environment.

Many PTs will require up to 120 minutes in which to administer. Additional time might be necessary for prework or group work, as required by a particular task.

References

Center for Universal Design (CUD). (1997.) About UD: Universal Design Principles. http://www.design.ncsu.edu/cud/about_ud/udprincipleshtmlformat.html (accessed February 13, 2009). Archived at http://www.webcitation.org/5eZBa9RhJ.

Council of Chief State School Officers (2002) *Models for Alignment Analysis and Assistance to States.* Occasional paper.

Council of Chief State School Officers (2006) *Aligning Assessment to Guide the Learning of All Students.* Reports of research project funded by the U.S. Department of Education and managed by the CCSSO SCASS Group on Technical Issues in Large-Scale Assessment.

Darling-Hammond, L. & Pecheone, R. (2010) *Developing an Internationally Comparable Balanced Assessment System That Supports High-quality Learning.* Princeton, NJ: Educational Testing Service. Retrieved from: http://www.k12center.org/publications.html.

DeMauro, G.E. (2004) *Test Alignment Considerations for the Meaning of Testing.* Paper presented at the CCSSO Annual Conference on Large Scale Assessment, Boston.

Hansen, E.G. & Mislevy, R.J. (2008). Design Patterns for Improving Accessibility for Test Takers with Disabilities. Princeton, NJ, ETS Research Report No. RR-08-49

Martone, A. and Sireci, S. (2009) *Evaluating Alignment Between Curriculum, Assessment and Instruction*. Review of Educational Research, 79:4, 1332-1361.

Mislevy, R. J. (1995). Test *Theory Reconceived.* White paper based on an invited address to the meeting of the National Council of Measurement in Education, Atlanta, GA, April 1993.

Mislevy, R.J. & Haertel, G. (2006). Implications for *E*vidence-centered Design for *Educational Assessment.* Educational Measurement: Issues and Practice, 4, 6–20.

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2003). *On the Structure of Educational Assessments.* Measurement: Interdisciplinary Research and Perspectives, 1, 3-67. National Governors Association Center for Best Practices & Council of Chief State School Officers, (2010.) *Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects.* Washington, DC.

National Research Council/NRC. (2001). Knowing *What Students Know*: The science and design of *Educational Assessment.* Committee on the Foundations of Assessment. J. Pelligrino, N. Chudowsky, & R. Glaser (Eds.), Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Quellmalz, E. S. & Moody, Mark. (2004). Models for Multi-level *State Science Assessment* Systems. Report commissioned by the National Research Council Committee on Test Design for K-12 Science Achievement.
Rose, D., and A. Meyer. (2000) Universal design for learning, associate editor column. *Journal of Special Education Technology* 15 (1): 66-67.

Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting Performance Standards For Student Achievement.* Report of the NAE Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels. National Academy of Education.
.

Webb, N.L. (1997) *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*. Research Monograph No. 6, National Institute for Science Education at the University of Wisconsin – Madison and the Council of Chief State School Officers, Washington, DC.

Webb, N.L., Horton, M., and O'Neal, S. (2002) *An Analysis of the Alignment Between Language Arts Standards and Assessments for Four States*. Paper presented at the American Educational Research Association Annual Meeting, New Orleans.

Webb, N.L. (2002) *An Analysis of the Alignment Between Mathematics Standards and Assessments for Three States*. Paper presented at the American Educational Research Association Annual Meeting, New Orleans.

Zhang, T., Haertel, G., Javitz, H., Mislevy, R., Murray, E., & Wasson, J. (2009). A *Design Pattern* for a *Spelling Bee Assessment* for Students with Disabilities. A paper presented at the annual conference of the American Psychological Association, Montreal, Canada.

Zucker, S., Sassman, C., Case, B. (2004) *Cognitive Labs.* Technical Report, Pearson, Inc., Iowa City, Iowa.