Making Good Use of New Assessments:

Interpreting and Using Scores

From the Smarter Balanced Assessment Consortium

Linda Darling-Hammond

Edward Haertel

James Pellegrino

With the Assistance of Soung Bae[1]

March, 2015[2]

[1] The authors gratefully acknowledge the very helpful review and feedback provided by Randy Bennett of the Educational Testing Service.
[2] This paper was commissioned by the Smarter Balanced Assessment Consortium.

Introduction

New assessments are currently being adopted across the nation in response to policies intended to ensure students' readiness for college and careers when they leave high school. The value of these new assessments will be maximized with proper interpretation and use, as they inform instructional plans for individual students, classroom practices, curriculum designs, professional development activities, and other policies.

This document outlines some of the key principles and information that should guide users of the new Smarter Balanced Assessments, based on psychometric standards for general test use and on specific features of these assessments. The suite of assessments offered by Smarter Balanced offers many advances that can help to strengthen teaching and learning. Like any other tests, they must be properly used to achieve these benefits. This paper offers guidance about how to understand and best employ the assessments and the scores they produce.

The Smarter Balanced Assessments

In 2009, the Council of Chief State School Officers and the National Governors Association Center for Best Practices came together to coordinate a state-led effort to develop the Common Core State Standards (CCSS). ("About the Standards," 2014). The goal of the collaboration was to establish clear and consistent education standards in mathematics and English language arts that would help prepare all students for success in college and careers. The CCSS define what a student should study and learn across a set of skill areas described in learning progressions and grade level expectations. Currently, 43 states, in addition to the District of Columbia, four U.S. territories, and the Department of Defense Education Activity, have voluntarily adopted the CCSS ("Standards in Your State," 2014). Most states have plans to fully implement the standards by the spring of 2015.

The adoption and implementation of the CCSS will require next-generation assessments. As the 2013 report of the *Gordon Commission on the Future of Assessment in Education* noted:

> To be helpful in achieving the learning goals laid out in the Common Core, assessments must fully represent the competencies that the increasingly complex and changing world demands. The best assessments can accelerate the acquisition of these competencies if they guide the actions of teachers and enable students to gauge their progress. To do so, the tasks and activities in the assessments must be models worthy of the attention and energy of teachers and students…. (p. 7).

The Smarter Balanced Assessment Consortium (Smarter Balanced or SBAC) is one of two state-led consortia[1] that are working to develop systems of assessments aligned to the CCSS. The goal of both consortia's efforts is to provide meaningful feedback and actionable information to ensure that all students are progressing toward attaining the knowledge and skills needed for postsecondary success.

The Smarter Balanced assessment system is geared towards ensuring that "all students leave high school prepared for post-secondary success in college or a career through increased student learning and improved teaching" (Smarter Balanced, 2013, p. 9). Toward that end, the Smarter Balanced assessments include both summative assessments that enable comparable reporting across states and interim assessments that can inform instruction during the school year. The system also includes formative tools to support teaching that are available in a digital library of professional development materials, instructional resources, and tools aligned to the CCSS and to the Smarter Balanced claims. The digital library is designed to help educators address key learning challenges and differentiate instruction as they implement the new standards.

The summative and interim assessments incorporate new item types that go beyond multiple-choice questions to include extended-response and technology-enhanced items, as well as performance tasks that allow students to demonstrate critical-thinking and problem solving skills.

The assessments also take advantage of new technologies to employ computer adaptive testing (CAT) which provides more precise information to educators about student achievement. Through CAT, the computer automatically adjusts the difficulty of questions based on the answers that the student provides. For example, questions answered correctly will generate more difficult questions whereas questions answered incorrectly will generate easier ones. "By adapting to the student as the assessment is taking place, these assessments present an individually tailored set of questions to each student and can quickly identify which skills students have mastered."[2]

The computer adaptive technology within the Smarter Balanced assessment primarily addresses content and skills that are defined by the standards associated with a student's enrolled grade level. After most of the test is completed, students who are consistently scoring at the top or bottom of the distribution may be directed toward items that are just above or below the grade level, allowing those students to be assessed with questions somewhat closer in difficulty to their current capabilities.

This time is in many ways an exciting one in education. The adoption of the CCSS has ushered in a new era of assessments designed to provide meaningful feedback to educators so that they can support students in acquiring the core knowledge and skills needed for success in college and careers, and for becoming lifelong learners. However, this is also a time of uncertainty. There is much that we do not know about how the implementation of the CCSS will transpire and how the new assessments will behave. Therefore, we must recognize the very significant challenges that are before us and be appropriately mindful of what we have yet to learn.

Goals of this Paper

There is much to be learned about how new standards and assessments will be implemented and to what effect. Fortunately, we are not starting from a blank slate. We are able to draw on previous research and expertise from the field of measurement and assessments to guide this current work. Since the mid-1990s, for example, the National Research Council's Board on Testing and Assessment (BOTA) has studied the properties, uses, and effects of tests on students and on school systems. In addition, the *Standards for Educational and Psychological Testing*,

which are jointly developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education,[3] provide us with a roadmap for the proper use of tests and for making sound inferences and decisions from the test data. As the Smarter Balanced assessment system is implemented, we can be guided by these and other well-grounded principles and lessons about how the assessments can best be interpreted and for what purposes they can best be used.

In what follows, we provide a brief discussion of the principles and processes that have shaped the Smarter Balanced Assessments, followed by a list of general principles of good test use gathered from experts in the field of measurement and assessments, the research literature, and the *Standards for Educational and Psychological Testing*. We then apply those lessons to the Smarter Balanced assessments and provide key recommendations for the sound use of these next-generation assessments, and for understanding and interpreting test data so that well-grounded decisions can be made about students, educators, and schools.

Principles and Processes Shaping the Smarter Balanced Assessment System

In 2010, the Smarter Balanced Assessment Consortium laid out its vision for an innovative assessment system intended to inform parents, students, teachers, and policymakers about student achievement in relation to the Common Core State Standards.[4] The components of this unified system are designed to work together to help ensure that all students meet the Consortium's overarching goal that all students leave high school prepared for postsecondary success in college or career through increased student learning and improved teaching. The Consortium has developed:

- A comprehensively designed assessment system that assesses the Common Core State Standards (CCSS) with selected response, open-ended, and performance items and tasks; balances summative, interim and formative components; and provides access to all students (e.g., students with disabilities, English learners).
- An online adaptive test administration with a secure item and performance task bank.
- A consolidated reporting system designed to enhance understanding of student progress toward college- and career-readiness for students, parents, educators, and policy makers.
- Supports for professional development focused on assessment literacy.

The Smarter Balanced assessments are the only ones available to schools in the United States that have all of these features – incorporating advanced item types that can measure higher–order thinking skills with open-ended performance tasks, utilizing computer adaptive technologies to measure achievement more precisely for a wider range of students, and integrating formative tools as well as interim and summative assessments in a cohesive, online system.

The system provides flexibility to member states as they decide the best approach for administering and reporting the Consortium's assessment in their own state, and for using the full range of tools to support learning.

The Consortium's work is guided by the following principles:

1. Assessments are grounded in a thoughtful, standards-based curriculum and are managed as part of an integrated system of standards, curriculum, assessment, instruction, and teacher development.
2. Assessments produce evidence of student performance on challenging tasks that evaluate the Common Core State Standards (CCSS).
3. Teachers are integrally involved in the development and scoring of assessments.
4. The development and implementation of the assessment system is a State-led effort with a transparent and inclusive governance structure.
5. Assessments are structured to continuously improve teaching and learning.
6. Assessment, reporting, and accountability systems provide useful information on multiple measures that is educative for all stakeholders.
7. Design and implementation strategies adhere to established professional standards.

Processes for Developing a High-Quality Assessment

The processes undertaken by the Consortium to enact these principles have sought to use best practices in test design and development, with deep engagement of content and assessment experts and wide participation of educators from K-12 and higher education, as well as members of the general public. No other assessment system has engaged as many educators and experts in as many aspects of test design, development, pilot testing, field-testing and validation as Smarter Balanced has done.

At the beginning of the grant, the Consortium assembled a Technical Advisory Committee (TAC) consisting of members who possess wide-ranging expertise in large-scale assessment and those who are on the cutting edge of issues related to the design and development of next generation comprehensive assessment systems. The TAC has met regularly over the grant period, providing technical advice and support on key decisions on all components of the assessment system. Over the course of the project, thousands of educators from schools and universities have also provided advice and support.

Evidence-Based Design. To meet the expectations of a high quality, next generation assessment, the Consortium made a commitment to employ an evidence-centered design (ECD) approach – a more rigorous approach to designing items and tasks. In this approach, it is not enough to state that an assessment is aligned to the CCSS or that it is measuring college and career readiness; rather, the assessment is built on a foundation that states the content and cognitive processes being measured, articulates how they are measured, and addresses the relative importance of the content and cognitive processes being measured.

Content Specifications. Using the CCSS, the Consortium assembled a team of experts in the fields of mathematics, mathematics education, English language arts/Literacy (ELA/L), and assessment along with primary authors of the CCSS to write content specifications for ELA/L and mathematics. This team worked together to create an initial draft of the content specifications in Summer 2011. In this document, the Consortium first established the assessment claims discussed above along with the evidence that the Consortium would need to collect in order to support each claim by grade level. The documents specify assessment targets and lay out accessibility strategies for English learners and students with disabilities to be considered in addressing each target. Consortium staff and the Consortium's Technical Advisory

Committee reviewed this initial draft. A revised version went through two rounds of public review that lasted nearly 30 days during which more than 200 individuals and organizations provided feedback on the content specifications. Using the public's feedback, the documents were revised and the claims were voted on by the Governing States.

Item and Task Development. Along with a test blueprint, the Consortium developed item and task specifications to ensure that the assessment items measure the assessments' claims. To do this, the specifications delineate the types of evidence that should be elicited for each claim within a grade level. Then, they provide explicit guidance on how to write items in order to elicit the desired evidence. Consortium items are created according to the principle of **universal design.** As the name suggests, the concept of universal design aims to create items that accurately measure the assessment target for as many students as possible. To facilitate the application of universal design principles, item writers are trained to consider the full range of students who may answer a question. A simple example of this is the use of vocabulary that will be known by all third-grade students versus only those third-grade students who play basketball. Almost all third-grade students are familiar with activities (e.g., recess) that happen during their school day, while only a subset of these students will be familiar with basketball terms like "double dribble," "layup," "zone defense," or "full-court press." Such terms are avoided in items developed using universal design principles.

Testing, Reviewing, and Revising Items. Using a set of item and task specifications derived from the content specifications, a small set of items was developed and administered in Fall 2012 during a small-scale trial. This provided the Consortium with its first opportunity to administer and score the new item types. During the small-scale trials, the Consortium conducted cognitive laboratories to better understand how students solve various types of items. A cognitive laboratory uses a think-aloud methodology in which students speak their thoughts while solving a test item. The Item and Task Specifications were again revised based on the findings of the cognitive laboratories. These specifications were used to develop items for the 2013 pilot test, and they were again revised based on the pilot test results.

A large-scale field test was administered to approximately 4.2 million students in over 16,500 schools across the 21 Governing States and the U.S. Virgin Islands in the spring of 2014. This field test allowed the Consortium to evaluate the performance of the more than 19,000 items and performance tasks in the item pool. Careful construction of the field test sample allowed the Consortium to conduct psychometric analyses on the data and to be assured that the results of these analyses were representative of the entire Consortium student population. In some cases, it was necessary to oversample some groups so that specific analyses could be conducted. For example, it was necessary to oversample Native American students in order to empirically study if test questions performed differently for this group than they did for other groups. The Consortium used the field test data to examine the items to understand which items performed well and which needed to be improved. This information will inform future item writing efforts.

Both before and after this field test, panels of educators reviewed all items, performance tasks, and item stimuli (e.g., reading passages) for accessibility, bias/sensitivity, and content. Four hundred four (404) mathematics educators from 14 states reviewed items and performance tasks; 262 ELA/L educators from 14 states reviewed items and performance tasks; and 95 educators

from 13 states reviewed the ELA/L stimuli. Items flagged for accessibility, bias/sensitivity, and/or content concerns were either revised or removed from the item pool.

Maximizing Accessibility. In order to provide every student with a positive and productive assessment experience and to generate results that are a fair and accurate estimate of each student's achievement, member states worked together to create an accessibility framework that includes universal tools, designated supports, and accommodations. These all yield valid scores when used in the manner specified by the Consortium's Guidelines. *Universal tools,* such as digital notepads and highlighters, are available to all students. Educators who are knowledgeable about a student's instructional needs may identify students who need *designated supports*, such as a separate test setting or masking of content that is distracting or not immediately needed, and can receive these in addition to the universal tools. Further *accommodations* are available to those students with documentation of the need through a formal plan (i.e., IEP or 504 plan). For example, American Sign Language may be used with listening items and print on demand can be used for students needing a paper-copy of the item. These aspects of the assessment were also evaluated in the pilot tests and field tests.

Engaging Educators in Developing Achievement Level Descriptors (ALDs). In October 2012, 30 K-12 educators and 21 representatives of higher education from two- and four-year colleges were convened to write ALDs. The K-12 educators were chosen to represent rural, suburban, and urban districts that had varying percentages of students receiving free and reduced lunch. These panelists represented all of the Governing States. For the Grade 11 claims, high school teachers and college faculty worked together to articulate the knowledge, skills, and processes that students would need to be considered ready for college and career. In addition to the ALDs, the Grade 11 panelists also reviewed and revised the Consortium's operational definition of College Content Readiness and Grade 11 policy framework. For the grades 3 – 8 claims, experienced educators created the ALDs.

Three rounds of review followed the workshop, including Smarter Balanced staff, committees, and more than 350 members of the public representing K-12 and higher education, who contributed to the wording of the final version. The operational definition of college content readiness, and the grade 11 policy framework was approved by the Governing States in April 2013.

Later, the Consortium involved thousands of constituents in setting achievement levels, using a process known as the "bookmark procedure." Close to 500 teachers, school leaders, higher education faculty, parents, business and community leaders met in person to review test questions and determined the threshold scores (i.e. cut scores) for four achievement levels for each grade and subject area. Representatives of each member state and educators with experience teaching English language learners, students with disabilities, and other traditionally under-represented students participated to help ensure that the achievement levels are fair and appropriate for all students. In addition, an online panel was open to educators, parents and other interested members of the community to provide input on the achievement levels. More than 9,500 people registered to participate in the online panel. A cross-grade review committee composed of 72 members of the in-person panels then took the results of the online and in-

person panels into account to develop recommendations that coherently aligned across grades and that reflected student progress from year to year.

As an additional step, Smarter Balanced engaged an external auditor, an Achievement Level Setting Advisory Panel and its standing Technical Advisory Committee to review the recommendations before they were presented to the states for approval. The auditor and both advisory panels certified that Smarter Balanced conducted a valid process that is consistent with best practice in the field.

As the Consortium approaches its first full administration of the new assessments in Spring 2015, it turns its attention to the ongoing validation and improvement of the new assessments. Having engaged in a development effort for an innovative unified assessment system deploying leading edge technical advances and taking advantage of exceptionally broad participation from the field, it is now time to consider how the assessments can be most thoughtfully implemented.

General Principles of Test Interpretation and Use

Educational assessments can offer valuable information to students, parents, educators, and policy makers regarding what students know and are able to do. When used appropriately, they can provide an objective and efficient way to gauge some aspects of student learning and achievement, and can inform the decision-making process about future instruction. All tests also have limitations. A single test cannot measure all the aspects of an individual's knowledge, skills, and abilities. And no test can measure learning perfectly. The following general principles of test score interpretation and use are generally accepted by measurement experts and are articulated in the newly revised *Standards for Educational and Psychological Testing*:

1) Tests are imprecise: Even the most well designed test has ***measurement error*** (AERA, APA, & NCME, 2014; NRC, 2007). Measurement error refers to the degree of imprecision or uncertainty in any assessment procedure. Measurement error occurs due to factors unrelated to student learning. For example, student performance on an assessment may be affected by mood, health, testing conditions, motivation, as well as uncertainty related to human scoring. Furthermore, the questions on a given test are only a sample of all the knowledge and skills that pertain to the subject being tested. If a different sample of questions had been chosen, or the questions had been posed in a different form, the student could have scored differently. Therefore, a test score is not an exact measure of a student's competencies since measurement error is inherent in all tests.

2) Tests provide only partial evidence about performance, thus they should be combined with other sources of evidence for decision making: In drawing any conclusion or making any decision, test scores should always be used in conjunction with ***multiple sources of evidence*** about performance (AERA, APA, & NCME, 2014; NRC, 2007). Consequential decisions about a student, educator, or a school should not be made only or primarily on the basis of a single test score. Because a test score is not perfect and only tells part of the story, other relevant information (e.g., student work samples, course grades, course taking record, teacher

observations, and other measures) should be included to place test scores in context and allow for a broader view of performance.

The extent and nature of evidence needed may depend on characteristics of the learner (e.g. age, prior schooling, native language, learning differences) as well as the interpretation to be made (e.g. next steps for instruction, program placement, readiness for a specific experience, etc.). A range of appropriate measures about an individual's competencies will enhance the validity of the overall interpretation of the test score and the appropriateness of decisions that rely in part on test data.

The **more consequential the test use, the stronger the evidence must be to support that use** (AERA, APA, & NCME, 2014; NRC, 2007). High stakes demand that a stronger body of additional supporting evidence is provided in order to "minimize errors of measurement or errors in classifying individuals into categories such as 'pass,' 'fail,' 'admit,' or 'reject'" (AERA, APA, & NCME, 2014, p. 188). When multiple sources of evidence agree, we can have greater confidence that the inferences we base on test scores are sound ones.

3) Validity depends on test design and use: A test is valid only when used with the intended population of test-takers for the specific purpose(s) and under the conditions (including prior preparation, motivation, and other administration conditions) for which it was designed and validated (AERA, APA, & NCME, 2014; NRC, 2007). Test validity refers to the extent to which inferences about individuals, based on their scores on a particular test, are defensible. When used as designed, test data can provide useful information. However, any test may function poorly or have unintended consequences if used outside of the specific purposes and populations for which it was designed and validated.

Test score interpretations or judgments are validated for specific purposes, and validity does not automatically transfer to new uses: each different purpose must be justified and validated in its own right. No assessment is valid for all possible purposes.

4) Opportunities to learn influence valid inferences as well as fairness: In educational contexts, valid inferences about student ability derived from tests depend on students having been provided opportunities to learn the tested material prior to the assessment being administered. The degree to which students are afforded high quality instruction and are supported to perform to their full potential impacts the degree to which test scores can appropriately support consequential decisions about their knowledge, skills, and abilities (NRC, 2007).

Interpretation and Use of Smarter Balanced Assessments

The general principles of test use and interpretation described above provide a foundation for using and interpreting the next-generation assessments that are being developed by Smarter Balanced and provide a framework for states to choose the best course in utilizing Smarter Balanced assessments. The Smarter Balanced assessments have been designed for specific purposes: to assess knowledge and skills included in the Common Core State Standards in

English language arts (ELA) and mathematics and, at the high school level, to support a determination about whether students are academically prepared to succeed in non-remedial courses in, or transferable to, a four-year college setting. In this section, we provide recommendations for putting the Smarter Balanced assessments to use in ways that conform to sound and ethical testing practices.

Interpreting Smarter Balanced Score Reporting

Smarter Balanced will report overall scores in ELA and mathematics as well as scores for each of the claims the tests are designed to evaluate. Assessment Claims are broad evidence-based statements about what students know and can do as demonstrated by their performance on subsets of the assessments. At each grade level within mathematics and ELA/literacy, there is one overall claim encompassing the entire content area and additional specific content claims. Students will receive a score on each overall claim and scores for the specific content claims. These scores are developed from clusters of items in the CAT as well as a more extended performance task in each grade level.

	English language arts and literacy[5]	Mathematics[6]
Overall, Grades 3-8	Students can demonstrate progress toward college and career readiness in English language arts and literacy.	Students can demonstrate progress toward college and career readiness in mathematics.
Overall, Grade 11	Students can demonstrate college and career readiness in English language arts and literacy.	Students can demonstrate college and career readiness in mathematics.
Claim 1	**Reading:** Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.	**Concepts & Procedures:** Students can explain and apply mathematics concepts and interpret and carry out mathematics procedures with precision and fluency.
Claim 2	**Writing:** Students can produce effective and well-grounded writing for a range of purposes and audiences.	**Problem Solving:** Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies.
Claim 3	**Speaking and Listening:** Students can employ effective speaking and listening skills for a range of purposes and audiences.	**Communicating Reasoning:** Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.
Claim 4	**Research/Inquiry:** Students can engage in research and inquiry to investigate topics, and to analyze, integrate, and present information.	**Modeling and Data Analysis:** Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.

Because specific content claims are each based on less evidence (i.e., fewer test items) than overall claims, they are necessarily less reliable. Thus, while claim-level scores are intended to be useful for guiding formative follow-up and instruction, overall scores will be more appropriate than claim scores for most purposes.

Interpreting and Using Scores

Smarter Balanced will report results in terms of both ***scale scores*** and ***achievement level descriptors*** for each tested student. ***Scale scores*** are the basic unit of reporting. A scaled score is derived from a total number of obtained score points that is statistically adjusted and converted into a consistent, standardized scale that permits direct and fair comparisons of scores from different forms of a test either within the same administration year or across years (Tan & Michel, 2011).

In the Smarter Balanced assessments, these scores are represented on a continuous vertical scale (from 2000 to 3000) that increases across grade levels. A "vertical scale" is one designed to support inferences concerning growth across years. In other words, with a vertical scale it should be possible, for example, to subtract a student's score on a third grade test from that same student's score on a fourth-grade test the following year to measure growth. Ideally, scale scores can be used to illustrate students' current levels of achievement and their growth over time. When aggregated, these scores can also describe school- or district-level changes in performance on the tests and can measure gaps in achievement among different groups of students. As we describe in a later section, however, growth metrics for Smarter Balanced assessments will need to be interpreted with considerable caution until further validation studies are completed.

Smarter Balanced has also developed a set of initial, policy ***achievement level descriptors*** (ALDs) for English language arts/Literacy (ELA/Literacy) and mathematics that are aligned with the Common Core State Standards (CCSS) and the Smarter Balanced assessment claims. The purpose of these descriptors is to enable states to comply with federal law and to specify, in content terms, the knowledge and skills that students display at four levels of achievement (i.e., Level 1, Level 2, Level 3, and Level 4) and to provide context, for those who are not familiar with the scale scores, to help them understand the meaning of the scores.

The Achievement Level Descriptors were developed based on the feedback of reviewers who engaged in a validation process including examination of the Common Core State Standards in each content area and items on representative forms of the Smarter Balanced assessments. As described earlier, the cut scores were recommended by panels of judges, including educators and curriculum experts, who used standard setting procedures to determine the regions on the score scale associated with demonstrating performance at a given achievement level. This very extensive process was thorough and professionally managed, with much wider engagement than the standard-setting process conducted for any assessment currently in use.

While this process supports a high level of confidence in the results, as the National Academy of Education (2009) has noted, the procedures for setting cut scores on any test is fundamentally judgmental. Even though systematic methods are employed to arrive at those decisions, there is

no "true" proficiency standard that can be objectively defined. To ensure that users avoid investing more meaning in those categories than is scientifically justified, these cut scores should be considered approximations of student abilities. Additional data will be collected over time to validate the achievement level descriptors in relation to the actual success rates of students in subsequent grade levels and when they enter college and careers. This long-term research will add to our knowledge about the meaning of the scores and descriptors.

While the achievement levels and their corresponding labels are a reporting feature that is federally required under the No Child Left Behind Act, and have thus become familiar to many educators and parents, such descriptors on any test should not be viewed as predicting students' futures, and should be used in the context of the multiple sources of information that we have about students and schools.

The Achievement Level Descriptors may aid in interpreting scores; however, they will be less precise than scale scores for describing student gains over time or changes in achievement gaps among groups, since they do not reveal changes of student scores within the bands defined by the achievement levels. Scale scores (rather than a label such as "percent proficient") will be more accurate for measuring year-to-year growth. Furthermore, there is not a critical shift in student knowledge or understanding that occurs at a single cut score point. Thus, the achievement levels should be understood as representing approximations of levels at which students demonstrate mastery of a set of concepts and skills. The scale scores just above and below the cut score for a given achievement level could be viewed as within the same general band of performance.

Interpreting Achievement Levels

The Achievement Level Descriptors are intended to describe the extent to which students have mastered the intended content for that grade level and are ready for subsequent elements of the curriculum. Although no specific Achievement level descriptor (i.e. a "2" or a "3") equates directly to expectations for "on grade-level" performance, they represent differing levels of performance and readiness for subsequent content for students within a grade level.

At the high school level, Achievement Level Descriptors are linked to an operational definition of ***college content-readiness*** to inform score interpretation for high schools and colleges. In particular, a score at or above "Level 3" in 11th grade is meant to suggest conditional evidence of readiness for entry-level, transferable, credit-bearing college courses, assuming the successful completion of senior year coursework. Since college readiness encompasses a wide array of knowledge, skills, and dispositions, only some of which can be measured by the Smarter Balanced assessments, "college readiness" in this context is defined as "content-readiness" in the core areas of ELA/Literacy and mathematics. In other words, the Smarter Balanced Assessments were designed to measure mastery of a set of standards which collectively represent a common agreement about the ELA and mathematics content stakeholder groups think students need to know and how they should be able to use it in order to be ready for college and careers requiring college.

High schools may combine scores at 11th grade with additional data (courses completed, grades, portfolios, performance assessments) to determine appropriate courses of study and supports for

students in the 12th grade to enhance a student's academic achievement and college readiness. Similarly, as colleges interpret scores on Smarter Balanced assessments, they are encouraged to evaluate additional data (courses completed, grades, portfolios, performance assessments) to determine advisement and placement in developmental or credit-bearing courses.

Currently, Smarter Balanced does not yet have a parallel operational definition and framework for ***career readiness***, thus scores do not support specific inferences about readiness for a range of careers and other postsecondary options. To address this, Smarter Balanced is convening a design group comprised of experts in career technical education and measurement to plan how Smarter Balanced can create a framework that does support inferences about career readiness and other post-secondary options.

Using Smarter Balanced Test Scores for Valid Purposes

Validity is a function of the uses to which we put tests, not the tests themselves. As is true for all tests, stakeholders are advised against using the Smarter Balanced test data for purposes outside of their intended use. The Smarter Balanced Summative assessments were designed to assess students' achievement and progress towards college and career readiness in English language arts and literacy, and mathematics. When used for that specific purpose, valid interpretations of the Smarter Balanced test scores regarding whether students are content-ready would be appropriate and defensible. Smarter Balanced has developed interim and formative resources to support additional purposes such as improving teaching and learning.

The Smarter Balanced assessments were not designed or validated for purposes such as assessing whether students should be promoted to the next grade or whether a student has demonstrated the competencies needed to graduate from high school. Hence, the results of the Smarter Balanced assessment would need to be specifically validated to evaluate whether they would offer defensible information for making those kinds of decisions before they could be considered for such use (Heubert & Hauser, 1999; NRC, 2007).

Interpreting Scores in Early Years of Implementation

Special challenges are introduced when new Content Standards demand new curriculum sequences and new pedagogical strategies, as the Common Core State Standards may in many states. These challenges will have an impact on student performance. As a consequence, caution and common sense must be practiced when interpreting the scores, especially in the early phases of implementation, as states are transitioning from their previous curriculum and tests to the Common Core State Standards and to the Smarter Balanced assessments. These same challenges would occur with the choice of any test to measure new standards. Because the standards for each grade level build on learning at prior grade levels, students' instructional history with CCSS-aligned curriculum sequences and pedagogical strategies may also affect their performance. In the early years of implementation, this may be an especially important consideration for students at higher grade levels.

When new content standards are assessed, the summative assessment scores will reflect both the degree to which the content standards are well-implemented in a school and the degree to which students have learned them. States will benefit from a well-planned approach to evaluating school practices as well as student learning, using multiple measures that can help shed light on the factors influencing test scores. States and districts may want to support schools to engage in a continuous improvement process to assess how student learning can be increased by modifying aspects of curriculum and teaching that affect implementation of adopted Content Standards.

As states plan their use of the test scores, they should consider the degree to which implementation of the standards is systemic prior to attributing the cause of low scores to students. In addition, states and school-based educators should incorporate multiple sources of evidence that reinforce one another in characterizing the achievement of a student – such as course grades, courses taken, samples of student work -- to enhance the validity of the overall interpretation of the test scores and make the decisions derived from the test data more defensible.

The same caution should be exercised with respect to the use of achievement level descriptors. If school staff members choose to use these descriptors to interpret test scores, they should be careful to avoid using negative descriptors to label children (e.g. "below basic" or "off-track" student). Substantial research strongly suggests that negative descriptors that children interpret as evidence of their ability can exert powerful influences on student behavior, learning, and demonstrations of accomplishment independent of actual knowledge or skill (Holme et al., 2010; Schmader et al., 2008).

Finally, while Smarter Balanced has completed extensive validation activities prior to launching the new assessments, it plans to hold itself to a higher standard that many tests have met in the past. Because the ultimate goal of the Smarter Balanced assessments is to measure students' postsecondary readiness, the Consortium plans to conduct predictive validity studies to examine how students fare after high school in relation to their scores on Smarter Balanced assessments. These data are still several years away from being available. When such data are available, users will know a great deal more about the meaning of scores at different points along the scale.

Generating and Interpreting Growth Scores

Although state-of-the-art methods were used to develop the Smarter Balanced vertical scale, inferences about growth are necessarily less accurate than inferences about status at one point in time, because the growth score on any test is influenced by measurement error in each of two annual tests. For all tests currently used by states in the U.S. for federal accountability purposes, this inaccuracy is exacerbated by the fact that, by federal requirement, the test items encountered by each student are primarily restricted to the grade level standards, even if the accurate achievement of the student would be better measured with items that are above or below the grade level standards. Because of its computer adaptive technology, within these constraints, Smarter Balanced tests are designed to provide a more accurate estimate of student achievement than other tests that lack the adaptive element.

Smarter Balanced scores for individual students (scale scores and/or achievement levels) will no doubt be used to create growth metrics for various kinds of analyses and studies. The limitations on this kind of measure for any test should be borne in mind: First, regardless of how well designed a test is for students within a given grade-level, there is inevitably differential sensitivity (information value) along the scale, with less precision and greater error in some parts of the scale than others. This problem of differential sensitivity along the score scale is generally less severe for adaptive tests like those provided by Smarter Balanced, although still present to some degree.

If states deploy growth models to examine student gains over time, they should consider the accuracy of the growth scores as part of their process for determining any uses and interpretations of the scores. The accuracy should be evaluated for low and high performing students as well as students in subgroups such as English language learners and students with disabilities.

Although Smarter Balanced uses computer adaptive testing, and although test scores will be arrayed on a vertical scale, the tests will not measure the full continuum of achievement along which students actually fall. As noted above, Federal rules require that students be tested with items associated with grade level content and skills and most students will encounter only those items. Based on their performance on the bulk of the test, the CAT will direct some students to items just above or below grade level. However, the test will not measure the actual skill levels of students whose achievement is far above or far below their grade levels, since they will not encounter many items that accurately measure their actual levels of achievement. Thus, while the test will accurately describe the student's knowledge of grade level and near–grade level content, the test will not be as sensitive a measure of growth for these students.

As an example, an 8th grade math teacher whose students are functioning with 4th grade skills will have to teach requisite computational skills while also seeking to teach algebraic concepts. Given the design parameters associated with Smarter Balanced assessments, gains that students make in these foundational computational skills will not be specifically measured on the tests, because students will not encounter items that would evaluate their growth of these specific skills. The growth metric that could be produced from Smarter Balanced tests will only provide information about the extent to which students' scores changed within the region of their grade level, not a full measure of their starting and ending skills at two points in time.

It is important to note that Smarter Balanced makes available an interim assessment that can be deployed at any grade level for formative purposes. Teachers can use this as an additional source of information about performance and growth across the full spectrum of student achievement to guide their teaching.

Finally, nonrandom decisions about student placements (e.g. through tracking or other nonrandom assignment across classrooms) and nonrandom student enrollment across schools (e.g. through school choice and the existence of economic and racial/ethnic segregation) may also systematically bias growth metrics. For these reasons, growth measures produced by any test should be interpreted with caution and used as part of a set of multiple measures that provide additional information about student progress.

The challenges are greater when analysts seek to make causal inferences about the reasons for measured growth in student test scores. Value-added methodologies (VAMs) that seek to draw inferences about the causes of student growth, while useful for large-scale research studies, have proven unreliable and error-prone when used for drawing inferences about the effects of individual teachers (Baker et al., 2010; McCaffrey et al., 2005; Haertel, 2013). One challenge is the impossibility of fully disentangling individual educator effects from those of other teachers, specialists, tutors, curriculum, class size, school resources, school leadership, families, and student health and welfare. For these and other reasons, value-added ratings have been found to exhibit both instability from year to year and class to class and bias against teachers working especially with certain groups of students. States should exercise great caution in using value-added methodologies to draw inferences about the influences of individual teachers or schools on student achievement gains and should assemble a range of data from multiple sources to provide insights about what supports those gains.

Using Smarter Balanced Assessments to Improve Instruction

A critical starting point for the appropriate use of Smarter Balanced tests is to create a policy framework within which the tests can be well used for informing and improving the instructional program, which should be their primary use. We address these considerations in the next section.

Training for Assessment Literacy

Appropriate and meaningful interpretations of Smarter Balanced test scores depend on the assessment literacy of those who administer and make use of the test results. Stakeholders must be informed consumers with a solid understanding of what the test results really mean as well as their limitations. As a result, ongoing professional development that focuses on improving assessment literacy should be provided to students, parents, educators, and policy makers.

These groups must understand that test scores are imprecise; that they measure a subset of what is important for students to learn; that they should never be used as the sole input into consequential decisions; that while they reflect some of what has been learned, they do not identify the causes of that learning; and that they can have unintended consequences that must be protected against. Assessment literacy will enhance the validity of the inferences about an individual's competencies derived from test scores and increase the likelihood that educationally sound and legally defensible decisions are made.

Utilizing the Full Suite of the Smarter Balanced Assessment System

The Smarter Balanced system of assessments includes both summative assessments for accountability purposes and interim assessments for instructional purposes as well as a digital library of research-based, formative assessment tools and professional development materials aligned to the CCSS and the Smarter Balanced claims and assessment targets.

The various system components are designed for different needs. For example, the Smarter Balanced summative test is best used for making inferences about broad trends in student achievement and for describing students' achievement based on College and Career Ready content standards. Smarter Balanced interim assessments, which are aligned to the summative assessments, provide more actionable information to guide instruction and allow teachers to track student progress throughout the school year. The digital library of formative assessment tools and resources will further help teachers to identify instructional strategies that are helpful for teaching the standards.

The suite of Smarter Balanced tools should be used in tandem, which exemplifies one of the principles of sound test use: Test scores should be used in conjunction with multiple pieces of evidence to arrive at a more complete understanding of student learning and progress. As an example, if the Smarter Balanced interim test results suggest that students are struggling with skills in a particular area, then evidence from other indicators such as formative assessments, classroom observations, or analyses of student work can be collected and analyzed to see if that suggestion can be confirmed. This other information can inform more fine-grained interpretation of students' understanding and guide teaching decisions.

Use Learning Progressions to Guide Instruction

Upon gathering diagnostic information through the use of the Smarter Balanced interim assessments and other relevant data, educators will be aided by focusing on the learning progressions in the CCSS to guide instruction as well as the grade level content standards. The learning progressions provide a general mapping of how knowledge, skills, and understandings are typically developed over time and they are reflected in the CCSS as the "skills and content that advance in difficulty from one grade to the next and guide the unfolding of curriculum and instruction over time" ("Content Specifications for the Summative Assessment," 2011, p. 17).

This recommendation is particularly apt given the need to balance meeting students where they are and the policy requirement that all students be held to a common set of standards. Holding all students to a common set of standards creates an impossible dilemma for teachers whose students enter the classroom several grade levels behind – the struggle becomes how simultaneously to meet both the system's expectations for teaching grade level content and the students' instructional needs for earlier content they have not yet mastered. The learning progressions provide a way to focus on essential or core concepts and processes that connect learning across grades and critically contribute to the construction of more mature understandings later on (Hess, 2008). In addition, learning progressions can reveal gaps in the key core concepts, which allow teachers to tailor instruction to specific learning needs. The learning progressions also can serve as a means for tracking student progress within and across grades to ensure that skills and understanding of concepts are deepened over time.

Conclusion

Over time, as educators and students develop greater understanding of and experience with the CCSS, improvements in student performance and changes in growth from year-to-year on the

Smarter Balanced assessments can be expected. Results from predictive validation studies will also place the interpretation of scores in a more meaningful perspective. In the meantime, users must be mindful of the ways in which test scores are interpreted and resist the temptation to make unwarranted decisions from them, as they should with any test.

The general principles of sound measurement, test validation, and test use remind us that no single test is a perfect measure of what an individual knows and is able to do. The questions on a given test are only a sample from the larger domain of knowledge and skills that students are expected to acquire. In addition, all tests, even well-designed tests, are subject to measurement error. Therefore, multiple sources of evidence must be used in order to lead to more valid decision-making that includes, but that does not rely solely upon, test score data. Moreover, valid inferences derived from test scores depend on the opportunities to learn provided to students prior to the test being administered. Without sufficient opportunities to engage with quality curricula and instruction, appropriate, accurate inferences about student competencies are difficult and subject to error, even with well-designed assessments.

As the new standards are implemented, their most important uses should be as signals to educators and students about the kinds of learning and instruction that are valued in the classroom. Assessment data should be used for information in evaluating progress and guiding ongoing curriculum and professional development decisions to improve instruction. Scores should be combined with other evidence of student work and learning as a basis for judgments about what students know and can do. In order for all students to be college- and career-ready upon graduation from high school, curriculum materials, instructional activities, and various types of assessments should be used together to structure students' opportunities to learn, thereby enabling them to think critically, problem solve, and demonstrate the 21st century skills needed for postsecondary success.

Standards for Educational and Psychological Testing*

Standard 1. Validity

1.0: Clear articulation of each intended test score interpretation for a specified use should be set forth and appropriate validity evidence in support of each intended interpretation should be provided.

1.1: The test developer should set forth clearly how test scores are intended to be interpreted and consequently used.

1.2: A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation.

1.3: If validity for some common or likely interpretation for a given use has not been evaluated, or if such an interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations.

1.4: If a test score is interpreted for a given use in a way that has not been validated, it is

incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence, if necessary.

Standard 2. Reliability / Precision and Errors of Measurement

2.0: Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.

Standard 3. Fairness in Testing

3.0: All steps in the testing process, including test design, validation, development, administration, and scoring procedures should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.

3.18: In testing individuals for diagnostic and/or special program placement purposes, test users should not use test scores as the sole indicators to characterize an in individual's functioning, competence, attitudes, and/or predispositions. Instead, multiple sources of information should be used, alternative explanations for test performance should be considered, and the professional judgment of someone familiar with the test should be brought to bear on the decision.

3.19: In settings where the same authority is responsible for both provision of curriculum and high-stakes decisions based on testing of examinees' curriculum mastery, examinees should not suffer permanent negative consequences if evidence indicates that they have not had the opportunity to learn the test content.

Standard 4. Test Design and Development

4.0: Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

Standard 5. Scores, Scales, Norms, Score Linking, and Cut Scores

5.0: Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.

5.3: If there is a sound reason to believe that specific misinterpretations of a score scale are likely, test users should be explicitly cautioned.

5.23: When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

Standard 6. Test Administration, Scoring, Reporting, and Interpretation

6.0: To support useful interpretations of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation, Those responsible for administering, scoring, reporting, and interpreting should have the sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any materials errors should be documented and, if possible, corrected.

Standard 7. Supporting Documentation for Tests

7.0: Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores.

7.1: The rationale for a test, recommended uses of the test, and information that assists in score interpretation should be documented. When particular misuses for a test can be reasonable anticipated, cautions against such misuses should be specified.

Standard 8. The Rights and Responsibilities of Test Takers

8.0: Test takers have the right to adequate information to help them properly prepare for a test so that the test results accurately reflect their standing on the construct being assessed and lead to fair and accurate score interpretations. They also have the right to protection of their personally identifiable score results from unauthorized access, use, or disclosure. Further, test takers have the responsibility to represent themselves accurately in the testing process and to respect copyright in test materials.

Standard 9. The Rights and Responsibilities of Test Users

9.0: Test users are responsible for knowing the validity evidence in support of the intended interpretations of scores on tests that they use, from test selections through the use of scores, as well as common positive and negative consequences of test use. Test users also have a legal and ethical responsibility to protect the security of test content and the privacy of test takers and should provide pertinent and timely information to test takers and other test users with whom they share test scores.

9.1: Responsibility for test use should be assumed by or delegated to only those individuals who have the training, professional credentials, and/or experience necessary to handle this responsibility. All specials qualifications for test administration or interpretation specified in the test manual should be met.

9.2: Prior to the adoption and use of a published test, the test user should study and evaluate the materials provided by the test developer. Of particular importance are materials that summarize the test's purposes, specify the procedures for test administration, define the intended

population(s) of test taker, and discuss the score interpretations for which validity and reliability/precision data are available.

9.3: The test user should have a clear rationale for the intended uses for a test or evaluation procedure in terms of the validity of interpretations based on the scores and the contribution the scores make to the assessment and decision-making process.

9.4: When a test is to be used for a purpose for which little or no validity evidence is available, the user is responsible for documenting the rationale for the selection of the test and obtaining evidence of the reliability/precision of the test scores and the validity of the interpretations supporting the use of the scores for this purpose.

9.5: Test users should be alert to the possibility of scoring errors and should take appropriate actions when errors are suspected.

9.6: Test users should be alert to potential misinterpretations of test scores; they should take steps to minimize or avoid foreseeable misinterpretations and inappropriate uses of test scores.

9.7: Test users should verify periodically that their interpretations of test data continue to be appropriate, given any significant changes in the population of test takers, the mode(s) or test administration, or the purposes in testing.

9.8: When test results are released to the public or to policy makers, those responsible for the release should provide and explain any supplemental information that will minimize the possible misinterpretations of the data.

Standard 12. Educational Testing and Assessment

12.1: When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described by those mandate the tests. It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test developer and/or user.

12.10: In educational settings, a decision or characterization that will have major impact on a student should take into consideration not just scores from a single test but other relevant information.

12.17: In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of the differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

*These excerpts are drawn from the *Standards for Educational and Psychological Testing,* prepared by the American Educational Research Association, American Psychological

Association, and National Council on Measurement in Education (Washington, DC: American Educational Research Association, 2014). Note that Standard 10 concerns psychological testing and Standard 11 concerns workplace testing; thus they are not reproduced here.

References

About the Standards. (2014). In Common Core State Standards Initiative. Retrieved on December 11, 2014, from http://www.corestandards.org/about-the-standards/

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, D.C.: Authors.

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Retrieved on December 15, 2014, from http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Au, W. (2007). High stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258–267.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, D.C.: Economic Policy Institute.

Computer Adaptive Testing. (n.d.). In Smarter Balanced Assessment Consortium. Retrieved on December 11, 2014, from http://www.smarterbalanced.org/smarter-balanced-assessments/computer-adaptive-testing/

Content Specifications for the Summative Assessment of the Common Core State Standards for English Language Arts and Literacy. (2011). Retrieved on December 15, 2014, from http://www.ode.state.or.us/wma/teachlearn/commoncore/sbac_ela_literacycontentspecifications.pdf

Gordon Commission on the Future of Assessment in Education (2013). *Policy report.* Available: http://www.gordoncommission.org/publications_reports.html.

Haertel, E. (2013). Reliability and Validity of Inferences About Teachers Based on Student Test Scores. Princeton, NJ: Educational Testing Service.

Hess, K. (2008). *Developing and using learning progressions as a schema for measuring progress*. Retrieved on December 15, 2014, from http://www.nciea.org/publications/CCSSO2_KH08.pdf

Heubert, J.P. & Hauser, R.M. (Eds.) (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation.* National Research Council. Washington, DC: National Academy Press.

Holme, J.J., Richards, M.P., Jimerson, J.B., & Cohen, R.W. (2010). Assessing the Effects of High School Exit Examinations. *Review of Educational Research, 80* (4): 476-526.

Koretz, D. (2008). *Measuring up: what educational testing really tells us*. Cambridge, MA: Harvard University Press.

McCaffrey, D.F., Koretz, D., Lockwood, J.R., & Hamilton, L.S. (2005). *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica: RAND Corporation.

National Academy of Education. (2009). *Standards, assessments, and accountability*. Washington, D.C.: Author.

National Research Council. (2007). *Lessons learned about testing: Ten years of work at the National Research Council*. Washington, D.C.: Author.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115* (2), 336-356. http://dx.doi.org/10.1037/0033-295X.115.2.336

Smarter Balanced Assessments. (n.d.). In Smarter Balanced Assessment Consortium. Retrieved on December 11, 2014, from http://www.smarterbalanced.org/smarter-balanced-assessments/

Smarter Balanced Assessment Consortium. (2013). *Content specifications for the summative assessment of the Common Core State Standards for mathematics*. Retrieved on December 11, 2014, from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Math-Content-Specifications.pdf

Smarter Balanced Assessment Consortium. *Interpretation and Use of Scores and Achievement Levels*. (2014). Retrieved on December 15, 2014, from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/11/Interpretation-and-Use-of-Scores.pdf

Standards in Your State. (2014). In Common Core State Standards Initiative. Retrieved on December 11, 2014, from http://www.corestandards.org/standards-in-your-state/

Tan, X., & Michel, R. (2011). *Why do standardized testing programs report scaled scores? Why not just report the raw or percent-correct scores?* Princeton, NJ: ETS. Retrieved on December 12, 2014, from http://www.ets.org/Media/Research/pdf/RD_Connections16.pdf

[1] The other consortium is the Partnership for Assessment of Readiness for College and Careers. Additional consortia are developing assessments for English learners and students with disabilities.
[2] "Computer Adaptive Testing," n.d., para.1.
[3] *Standards for Educational and Psychological Testing*, which are jointly developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, (2014)
[4] This section draws on the Smarter Balanced Assessment Consortium End of Grant Report to the U.S. Department of Education.

[5] See http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/05/ELA_Preliminary_-Blueprint-2014_04-30Final.pdf

[6] See http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/05/Math_Preliminary_-Blueprint-2014_04-30Final.pdf