Smarter Balanced Assessment Consortium:

Interim Assessment Technical Report for Educators

July 2017

Prepared for the Smarter Balanced Assessment Consortium

Submitted to:

Tony Alpert, Executive Director
Smarter Balanced Assessment Consortium
Patricia Reiss,
Senior Director, Systems Design
Matthew Schulz, Director, Psychometrics

Submitted by:

WestEd
Standards, Assessment, and Accountability Services

Deborah Sigman, Project Director
Kathryn Rhoades, Lead Developer

Table of Contents

Smarter Balanced Interim Technical Report for Educators

Introduction

Overview

The Smarter Balanced Assessment Consortium (Smarter Balanced) assessment system includes a set of balanced assessment components designed to meet the diverse student needs of its member states. This system provides valid, reliable, and fair assessment of deep disciplinary understanding, higher-order thinking skills, and rigorous college and career readiness standards. The system is based on the belief that assessment must support ongoing improvements in student instruction and learning experiences that lead to outcomes valued by all stakeholders. It is grounded in the strong foundational assessment policies and procedures of the Smarter Balanced member states, including supports and resources from institutions of higher education and workplace representatives. Smarter Balanced represents a high-quality, balanced, multistate assessment system aligned to the Common Core State Standards (CCSS) in English language arts/literacy (ELA/literacy) and mathematics. The intent of this report is to provide information about the Smarter Balanced interim assessments, a component of the larger Smarter Balanced assessment system, and evidence in support of their validity[1]. This report focuses on both Smarter Balanced interim assessment offerings: the Interim Comprehensive Assessments (ICAs) and the Interim Assessment Blocks (IABs).

This report provides information about the overall assessment system for context. The system includes:

- Summative assessments that determine students' progress toward college and career readiness in ELA/literacy and mathematics. The summative assessments are given at the end of the school year and consist of two parts: a computer adaptive test (CAT)[2] and a performance task[3]. These summative assessments incorporate a variety of item types, including technology-enhanced items, items that require a constructed response, and performance tasks. Items are deliberately designed to measure specific content. The assessments include writing at every grade and ask students to solve multistep, real-world problems in mathematics.
- Interim assessments that allow teachers to check students' progress at mastering specific concepts at strategic points throughout the school year, and that provide them with information that they can use to improve instruction and help students meet the challenge of college and career readiness standards. These tools are used at the discretion of schools and districts. There are two types of Smarter Balanced interim assessments: Interim Comprehensive Assessments (ICAs), which test the same content, and report scores on the same scale, as the summative assessments, and Interim Assessment Blocks (IABs), which focus on smaller sets of related concepts and provide more detailed information for instructional purposes. The interim assessments incorporate items that are developed using the same processes as are used to develop the items on the summative assessments. The interim assessments provide administration options that are more flexible than the summative assessment to assist educators in determining what students know and can do

[1] The degree to which each interpretation or use of a test score is supported by the accumulated evidence.
[2] A method of testing that actively adapts to the test taker's ability level during a computerized assessment.
[3] An assessment component that involves significant student interaction with stimulus materials and/or engagement in a problem solution, ultimately leading to an exhibition of the student's application of knowledge and skills.

in relation to the CCSS. In contrast to the summative assessments, these interim assessments are currently only available as fixed forms[4].

- The Smarter Balanced Digital Library is an online collection of high-quality instructional and professional learning resources contributed by educators, for educators. These resources have been developed to help educators implement a formative assessment[5] process to improve teaching and learning. Educators can use the resources to engage in professional learning communities, differentiate instruction for diverse learners, engage students in their own learning, improve assessment literacy, and design professional development opportunities. The Digital Library also incorporates features that provide educators with opportunities to comment on and rate resources and to share their expertise with colleagues across the country in online discussion forums.

Report Approach

While the intent of this report is to provide evidence in support of the validity[6] and value of the Smarter Balanced interim assessments, readers should recognize that demonstration of validity is an ongoing process. This report has been written with the understanding that the flexibility of the interim assessment system allows for educators to make local decisions about its best and most appropriate uses; therefore, validity is impacted by the inferences and decisions that educators will make based on their administration procedures.

States do not provide response data or scores from interim assessments to Smarter Balanced for analysis. Consequently, much of the evidence provided in this report focuses on the development of test items and on characteristics of test forms. Unlike the Smarter Balanced summative assessments, the interim assessments are not secure. Smarter Balanced member states retain flexibility regarding how to customize the interim assessment system so that it may best be used as part of their approaches to improve their local educational systems. The interim assessments may be administered in a standard manner, or teachers may use items or tasks from these assessments as a basis for classroom discussion and instruction or individual feedback. In addition, the interim assessments may be administered to the same students several times during the year.

Purposes of the Smarter Balanced Interim Assessments

Interim assessments, along with formative assessment strategies, offer instructionally useful information to teachers and students. They provide valid, reliable, and fair information about:

- Student progress toward mastery of the ELA/literacy and mathematics skills that are measured by the Smarter Balanced summative assessments;
- Student progress toward the mastery of skills measured in ELA/literacy and mathematics across all students and subgroups;

[4] All students receive the same questions, regardless of individual students' performance on past questions. Opposite of adaptive testing.
[5] Tests designed to provide feedback to teachers so that they can adjust instruction.

- Student performance at the claim[7] or cluster of assessment targets[8] level, so that teachers and administrators can monitor student progress throughout the year and adjust instruction accordingly;
- Teacher-moderated scoring of interim performance events as a professional development vehicle to enhance teacher capacity to evaluate student work aligned to the standards.

Overview of Report Chapters

Overview of report chapters.

CH#	Chapter Title
1	Validity
2	Test Fairness
3	Test Design
4	Test Administration
5	Reporting and Interpretation

The following text provides brief synopses of the content of each chapter of this report, and practical descriptions of the purpose of evidence in each chapter to direct further review and provide context for the interim assessments within the larger assessment system.

Chapter 1: Validity. Validity evidence is provided throughout this report. This chapter on validity provides information about test purposes and an overview of how interim assessment scores are appropriate for those purposes, including a statement of test purposes; valid score uses and interpretations; and an outline of validity evidence in this report.

This chapter provides information to answer the following questions:

- For what purposes were the interim assessments designed to be used?
- What evidence shows that test scores are appropriate for these uses?
- What are the intended test-score interpretations for specific uses?

Chapter 2: Test Fairness. Test fairness concerns whether score interpretations are valid and minimize construct-irrelevant variance[9] for all relevant subgroups. The evidence for test fairness of the interim assessments is based on the Smarter Balanced accessibility framework and on item development processes. All Smarter Balanced interim assessment items are fully accessible.

[7] The concept or characteristic that an assessment is designed to measure. Also referred to as a construct.
[8] Targets are the bridge between the content standards and the assessment evidence that supports the claim. They ensure sufficiency of evidence to justify each claim.
[9] The introduction of extraneous, uncontrolled variables that affect assessment outcomes.

This chapter provides information to answer the following questions:

- How can it be ensured that the interim assessments are fair for all students?
- How is fairness considered in developing test items and tasks?
- How are the tests administered so that each student can demonstrate his or her skills?

This chapter presents the Smarter Balanced Accessibility and Accommodations Framework, as well as information on bias and sensitivity[10] reviews conducted during item and task development and on differential item functioning (DIF)[11] analyses.

Chapter 3: Test Design. This chapter provides information pertaining to the content validity of the Smarter Balanced interim assessment system. It describes how tasks and items are structured to achieve desired domain[12] coverage; provides evidence that the assessments address the knowledge and skills that are required for college and career readiness; describes test structure (claims and targets) and its relationship to the CCSS, item and task development, and alignment[13] studies; and contains information about operational use[14] blueprints and test scoring methods.

This chapter provides information to answer the following questions:

- What is on the test?
- Is the test content consistent with stated test purposes?

Chapter 4: Test Administration. Test validity partially rests on the assumption that interim assessments are administered in a standardized manner. The Smarter Balanced interim assessments are administered on a large scale, in different policy and operational contexts, and may be administered at multiple points throughout the school year, at grades 3–8 and high school; individual assessments may be administered at any grade level. Thus, Smarter Balanced provides an administration manual template that member states may customize for specific use. Chapter 4 describes the customizable online Smarter Balanced Test Administration Manual (TAM).

This chapter provides information to answer the following questions:

- What are the test administration conditions that will best ensure that every student has been afforded the same chance for success?
- How can the test be administered to allow for accessibility for all students?
- Was the test administration secure or non-secure?
- Do test records show that the test was administered as intended?

---

[10] *Bias* is prejudice in favor of or against a subgroup. *Sensitivity* is awareness of the need to avoid bias in assessments.
[11] A statistical indicator of the extent to which different groups of test takers who are at the same ability level have different frequencies of correct responses, or, in some cases, different rates of choosing various item options.
[12] A group of related standards.
[13] The correspondence between student learning standards and test content.
[14] The actual use of a test to inform an interpretation, decision, or action, based in part or wholly on test scores.

Chapter 5: Reporting and Interpretation. This chapter provides examples of the Smarter Balanced suite of reports and interpretive information, along with an explanation of report elements. It also discusses intended uses of report information.

This chapter provides information to answer the following questions:

- What information do Smarter Balanced reports on the interim assessments contain?
- What do scores mean?
- How can teachers and parents best use the reports?

Chapter 1 – Validity

Introduction

Validity refers to the degree to which each interpretation or use of a test score is supported by the accumulated evidence (AERA, APA, & NCME, 2014; ETS, 2002). Validity is the central notion underlying the development, administration, and scoring of a test and the uses and interpretations of test scores. *Validation* is the process of gathering evidence to support each proposed score interpretation or use. The validation process does not rely on a single study or on collecting one type of evidence. Rather, validation involves multiple investigations and different kinds of supporting evidence (AERA et al., 2014; Cronbach, 1971; ETS, 2002; Kane, 2006). It begins with test design and is implicit throughout the assessment process, which includes development, field testing[15] and analyses of items, test scaling[16] and linking[17], scoring, and reporting. This chapter provides a framework for the validation of the Smarter Balanced interim assessments that largely overlaps the framework for the validation of the summative assessments.

Purposes of the Smarter Balanced Interim Assessment System

The Smarter Balanced assessment system includes two types of interim assessments, each with different purposes: the Interim Comprehensive Assessments (ICAs) and the Interim Assessment Blocks (IABs). The ICAs use the same assessment blueprints as the summative assessments and assess the same standards. When they are administered under standard conditions, the ICAs deliver a valid overall score and associated error of measurement and an indicator of performance at the claim level. Unlike the summative assessments, ICAs are fixed-form tests. The IABs focus on smaller sets of targets associated with an instructional block or unit. They are short fixed-form tests that can be used more flexibly than the ICAs to support teaching and learning. Importantly, items on the ICAs and IABs are not initially identified as items for the interim assessments, but are chosen from a general pool of items developed for both the summative and interim assessments.

Interim assessments can be used in a variety of ways. They can be administered under standard conditions, as described in the Smarter Balanced Test Administration Manual (TAM). They can also be administered repeatedly to a class or an individual student. In addition, they may be used as a basis for class discussion or feedback at the item level. Because information about the reliability of test scores applies only to the first time the test is administered under standard conditions, subsequent administrations and non-standard administrations, such as results from collaborating with another class or teacher, may alter the interpretation of results. The following interim assessment purposes apply only to the initial standard test administration.

The four main purposes of the interim assessments are to provide valid, reliable, and fair information about:

[15] A test administration that is used to check the adequacy of testing procedures and the statistical characteristics of new test items or new test forms. A field test is generally more extensive than a pilot test (administered to a sample of test takers to try out some aspect of the test or test items)
[16] The process of creating a number score (see scale and scale score in Appendix) to enhance test score interpretation, by placing scores from different tests or test forms on a common scale or by producing scale scores designed to support score interpretations.
[17] The process of relation scores on tests.

- Student progress toward mastery of the skills measured in ELA/literacy and mathematics by the Smarter Balanced summative assessments;
- Student performance at the claim or content cluster level[18], so that teachers and administrators can monitor student progress throughout the year and adjust instruction accordingly;
- Individual and group (e.g., school, district) performance in ELA/literacy and mathematics at the claim level, to determine whether students are adequately progressing; and
- Student progress toward the mastery of skills measured in ELA/literacy and mathematics across all students and subgroups of students.

Summary of Validity Argument

The core of the argument presented in this report is that the technical quality of the interim assessments supports the aforementioned purposes. The Common Core State Standards (CCSS) http://www.corestandards.org) are widely recognized content standards for college and career readiness in high school grades and for being on track for college and career readiness in lower grades (Conley et al. 2011. Content specifications and test blueprints show that the Smarter Balanced ICAs cover the breadth and depth of assessable standards (see Appendix for links to these documents). The assessments contain a variety of item types that allow for response processes designed to elicit a wide range of skills and knowledge. The IABs are designed to deliver information that, when combined with other information, is suitable for informing instructional decisions. IAB and ICA score reports indicate approaches for gaining further instructional information through classroom assessment and observation.

Smarter Balanced Scale Development

The Smarter Balanced vertical scale was constructed to provide measurement across grades, facilitating estimates of progress toward career and college readiness. The appropriateness of Smarter Balanced achievement standards as predictors of college and career readiness in grade 11 and of being on track for next-grade readiness in grades 3–8 was established by an extended achievement level setting process involving many K-12 educators from across the Smarter Balanced states. Participants from member states' higher education systems were also involved, to ensure that readiness criteria represented content knowledge and skills that are needed in college. Further information about these processes can be found in the Smarter Balanced 2014-15 Technical Report https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf, scoring specifications http://www.smarterbalanced.org/assessments/development/, and achievement level determination descriptionhttp://www.smarterbalanced.org/assessments/scores/

[18] Related content that can be measured by assessing similar skills

Validity Framework for Interim Assessments

The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) *Standards for Educational and Psychological Testing* (2014), hereafter referred to as "the *Standards*," describe a process of validation that consists of developing a convincing argument, based on empirical evidence, that the interpretations and actions based on test scores are sound:

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. (pp. 21–22)

The validity framework for the Smarter Balanced interim assessments corresponds to five sources of validity evidence that are described in the *Standards* (pp. 26–31). These five sources are:

- Evidence Based on Test Content
- Evidence Based on Internal Structure
- Evidence Based on Relations to Other Variables
- Evidence Based on Response Processes
- Evidence for Validity and Consequences of Testing

Because validity is an ongoing process with continuous addition of evidence from a variety of contributors, this report summarizes development and performance of the assessment itself, addressing test content, response processes, and internal structure. Other elements of validity evidence may come from supplemental research projects or third-party studies. Certain types of validity evidence are not available for teacher-administered non-secure ICAs and IABs.

As the *Standards* note, "[v]alidation is the joint responsibility of the test developer and the test user" (AERA et al., 2014, p. 13). Each Smarter Balanced member state and/or its local educational agencies determine how the interim assessments are used. Smarter Balanced provides information about test content and technical quality, and provides guidance to members on appropriate uses of interim assessment scores.

Table 1 shows the sources of validity evidence that are discussed in this report, in relation to each of the four purpose statements for validation of the interim assessments.

Table 1.1. Sources of validity evidence.

Purpose	Source of Validity Evidence for Interim Assessments				
	Test Content	Internal Structure	Relations to Other Variables	Response Processes	Testing Consequences
1. Provide valid, reliable, and fair information about student progress toward mastery of the skills measured in ELA/literacy and mathematics by the Smarter Balanced summative assessments.	✓	✓		✓	
2. Provide valid, reliable, and fair information about student performance at the content cluster level, so that teachers and administrators can monitor student progress throughout the year and adjust instruction accordingly.	✓	✓			✓
3. Provide valid, reliable, and fair information about individual and group performance in ELA/literacy and mathematics at the claim level, to determine whether students are adequately progressing.		✓	✓		✓
4. Provide valid, reliable, and fair information about student progress toward the mastery of skills measured in ELA/literacy and mathematics across all students and subgroups of students.	✓	✓	✓	✓	✓

Interim Assessment Purpose 1: Provide valid, reliable, and fair information about student progress toward mastery of the skills measured in ELA/literacy and mathematics by the Smarter Balanced summative assessments.

To support this purpose, validity evidence should confirm that the knowledge and skills being measured by the interim assessments cover the knowledge and skills measured on the summative assessments and that the scores from the interim assessments are on the same scale as those from the summative assessments. The ICAs cover the depth and breadth of the knowledge and skills measured on the summative assessments. The IABs are not comprehensive, but provide information about one or more assessment targets within each subject. As indicated in **Error! Reference source not found.**, the studies providing this evidence are primarily based on test content, internal structure, and response processes.

Validity Studies Based on Test Content. The content validity studies conducted for the summative assessments provide information that is relevant to the interim assessments. As previously noted, during the item development process, items were developed without being designated for use on the interim assessments or on the summative assessments. The ICA blueprint reflects the content coverage and proportions on the summative assessments. For the IABs, content experts designed blueprints around target groupings that they judged to be most likely to comprise an instructional unit. In combination with a teacher's knowledge and professional judgment, IAB reports can be a valuable component of a full picture of students' knowledge and skills.

Validity Studies Based on Internal Structure. Scores from the ICAs are on the same scale as those from the summative assessments, and ICA items meet the same measurement criteria as items on the summative assessments. The structure of ICAs follows that of the summative tests, with nested hierarchical relationships between claims and targets.

IAB designs are based on expected instructional groupings as shown in IAB blueprints. Figure 1.1 graphically displays this hierarchical relationship.

Figure 1.1. Item Development Hierarchy

Also within the realm of internal structure (see Table 1.1) is evidence regarding the reliability or measurement precision of scores from the interim assessments. A lower level of measurement precision in the interim assessments is tolerable because the stakes are lower, there are multiple assessments, these assessments supplement the summative assessments, and results are combined with additional information when used instructionally.

Smarter Balanced does not collect scoring data from interim assessments, so only the properties of test forms are analyzed. However, Smarter Balanced does provide resources, training, and validity papers for all interim items requiring hand-scoring.[19]

Validity Studies Based on Response Processes. This interim assessment purpose relates to skills measured on the summative assessments, so the validity studies based on response processes that were described for the summative assessments are relevant to the interim assessments, in that they can be used to confirm that the interim assessment items are measuring higher-order skills.[20]

Interim Assessment Purpose 2: Provide valid, reliable, and fair information about student performance at the content cluster level, so that teachers and administrators can monitor student progress throughout the year and adjust instruction accordingly.

As shown in **Error! Reference source not found.**, validity evidence to support this purpose of the interim assessments relies on studies of test content, internal structure, and testing consequences.

Validity Studies Based on Test Content. The Human Resources Research Organization (HumRRO) conducted an independent alignment study to gather evidence about the alignment of the Smarter Balanced summative assessments to the CCSS. The alignment analysis considered range of content, balance of content, and cognitive complexity. To determine these aspects of alignment, HumRRO conducted a series of workshops during which the participants reviewed the alignment among the Smarter Balanced summative assessment blueprints and the CCSS. The results of this alignment study pertain to the ICAs since the ICAs follow the same blueprints as the summative assessments. The alignment study report can be found at https://portal.smarterbalanced.org/library/en/smarter-balanced-assessment-consortium-alignment-study-report.pdf.

Validity Studies Based on Internal Structure. Information regarding the reliability and measurement error of cluster-level IAB score reporting can be found in Chapter 2 of the *Smarter Balanced Assessment Consortium 2014-15 and 2015-16 Technical Report*, available at https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf.

Interim Assessment Purpose 3: Provide valid, reliable, and fair information about individual and group performance in ELA/literacy and mathematics at the claim level, to determine whether students are adequately progressing.

As shown in **Error! Reference source not found.**, validity evidence to support this purpose of the interim assessments relies on studies of internal structure, relations to other variables and testing consequences.

Validity Studies Based on Internal Structure. This purpose statement is similar to Interim Assessment Purpose 2, and the studies described for that purpose are also relevant to this purpose.

[19] See Appendix for links to educator resources.

[20] For more information about the validity studies done on the 2015-16 summative assessments: http://www.smarterbalanced.org/assessments/development/

However, rather than a focus at the content cluster level, the focus of this purpose is on the claim level.

Validity Studies Based on Testing Consequences. Smarter Balanced does not collect interim assessment data from its members, so only analyses of properties of test forms can be conducted.

Interim Assessment Purpose 4: Provide valid, reliable, and fair information about student progress toward the mastery of skills measured in ELA/literacy and mathematics across all students and subgroups of students.

Educators will want to consider any consequences of the results of the interim assessments. Consequences will depend on how the results are intended to be used. Validity evidence in support of this purpose should come from all five sources. The validity studies based on test content that were described with respect to Interim Assessment Purposes 1 and 2 provide a starting point for equitable measurement across all students. Validity studies based on internal structure should report any estimates of reliability, measurement precision, decision consistency, or decision accuracy separately for all subgroups of students, and for students who take different variations of the interim assessments.

Conclusion for Interim Assessment Validity Results

Validation is an ongoing, essentially perpetual endeavor. This is particularly true for the many purposes typically given for assessments. Program requirements are often subject to change, and the populations that are assessed change over time. Nonetheless, at some point, decisions must be made regarding whether sufficient evidence exists to justify the use of a test for a particular purpose. The essential validity elements presented in this chapter constitute critical evidence "relevant to the technical quality of a testing system" (AERA et al., 2014, p. 22). This report describes how this evidence supports the use of the interim assessments as a tool to improve instruction and to help students meet the challenge of college and career readiness standards.

Chapter 2 – Test Fairness

Introduction

Smarter Balanced has designed its assessment system to provide all eligible students with fair assessments and equitable opportunities to participate in the assessments. Issues of test fairness apply to the entire assessment system, including both summative and interim assessments. Ensuring test fairness is a fundamental part of validity and is an important feature built into each step of the test development process, starting with test design and including item writing, test administration, and scoring. According to the *Standards*, fairness to all individuals in the intended population is an overriding and fundamental validity concern: "The central idea of fairness in testing is to identify and remove construct-irrelevant barriers[21] to maximal performance for any examinee" (AERA et al., 2014, p. 63).

The Smarter Balanced assessment system is designed to provide valid, reliable, and fair measures of student achievement based on the CCSS. The validity and fairness of these measures are influenced by a multitude of factors. Central among them are:

- clear definition of the construct—the knowledge, skills, and abilities—that is intended to be measured;
- development of items and tasks that are explicitly designed to assess the construct that is the target of measurement;
- delivery of items and tasks that enable students to demonstrate their achievement of the construct; and
- capture and scoring of responses to those items and tasks.

Smarter Balanced uses several processes to address fairness. The interim assessments use the same content specifications as the summative assessments and are fully accessible. This means that students have access to the same resources on the interim assessments that are available on the summative assessments. The *Smarter Balanced Content Specifications for the Summative Assessment of the CCSS for English Language Arts/Literacy* (Smarter Balanced, 2015a; https://portal.smarterbalanced.org/library/en/english-language-artsliteracy-content-specifications.pdf) and the *Smarter Balanced Content Specifications for the Summative Assessment of the CCSS for Mathematics* (Smarter Balanced, 2015b; https://portal.smarterbalanced.org/library/en/mathematics-content-specifications.pdf) define the knowledge, skills, and abilities to be assessed and their relationship to the CCSS. These documents describe the major constructs (identified as "claims"), within ELA/literacy and mathematics, for which evidence of student achievement is gathered and which forms the basis for reporting student performance. Each claim is accompanied by a set of assessment targets that provide more detail about ranges of content and Depth of Knowledge (DOK) [22]levels. The targets serve as the building blocks of test blueprints.

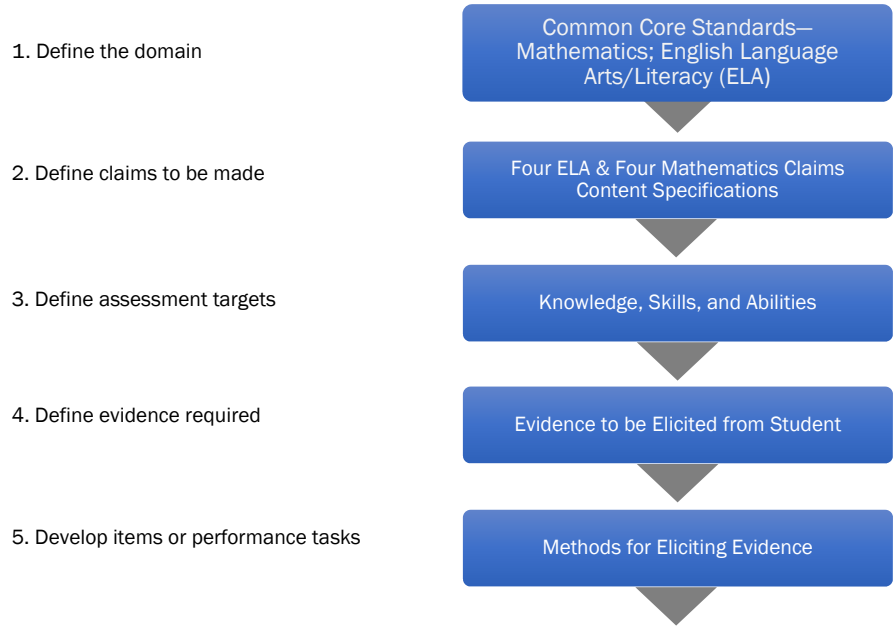
Figure 2.1 provides a graphical display of the evidence-centered design[23] process.

[21] Extraneous, uncontrolled variables that affect assessment outcomes.

[22] [22] A ranking of items and tasks (from 1–4) according to the complexity of thinking required to successfully complete them.

[23] Guiding approach for constructing SB assessments.

Figure 2.1. Concepts of evidence-centered design.



Much of the evidence presented in this chapter pertains to fairness to students during the testing process and to design elements and procedures that serve to minimize measurement bias. This chapter also addresses fairness in item and test design processes and the design of accessibility supports (e.g., universal tools, designated supports and accommodations) in content development.

Attention to Bias and Sensitivity in Test Development. According to the *Standards* (AERA et al., 2014), "bias" is "construct underrepresentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers and consequently the reliability/precision and validity of interpretations and uses of their test scores" (p. 216), and "sensitivity" refers to an awareness of the need to avoid explicit bias in assessment. Reviews of tests for bias and sensitivity help ensure that test items and stimuli are fair for various groups of test takers (AERA et al., 2014, p. 64).

The goal of fairness in assessment is to ensure that test materials are as free as possible from unnecessary barriers to the success of diverse groups of students. Fairness must be considered in all phases of test development and use. Smarter Balanced developed *Bias and Sensitivity Guidelines* (ETS, 2012; https://portal.smarterbalanced.org/library/en/v1.0/bias-and-sensitivity-guidelines.pdf) to help ensure that its assessments are fair for all groups of test takers, despite differences in characteristics including, but not limited to, disability status, ethnic group, gender, regional background, native language, race, religion, sexual orientation, and socioeconomic status. Smarter Balanced strongly relied on the Bias and Sensitivity Guidelines in the development and design phases of the assessments, including item-writer training, item writing, and review. Its focus and attention on bias and sensitivity are responsive to Chapter 3, Standard 3.2, of the *Standards*, which states: "Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests being affected by construct-irrelevant characteristics such

as linguistic, communicative, cognitive, cultural, physical or other characteristics" (AERA et al., 2014, p. 64).

According to these guidelines, unnecessary barriers can be reduced by following some fundamental rules

- measuring only knowledge or skills that are relevant to the intended construct;
- taking steps to avoid angering, offending, upsetting, or otherwise distracting test takers; and
- treating all groups of people with appropriate respect in test materials.

These rules help ensure that test content is both fair for test takers and acceptable to the many stakeholders and constituent groups within Smarter Balanced member organizations.

The Smarter Balanced Accessibility and Accommodations Framework

Smarter Balanced has built a framework of accessibility for all students, including, but not limited to, English learners (ELs), students with disabilities, and ELs with disabilities. Three resources—the item specifications for each grade band (available at http://www.smarterbalanced.org/assessments/development/), the *General Accessibility Guidelines* (Smarter Balanced, 2012; https://portal.smarterbalanced.org/library/en/general-accessibility-guidelines.pdf), and the Bias and Sensitivity Guidelines—are used to guide the development of Smarter Balanced assessments, items, and tasks and to ensure that they accurately measure the targeted constructs. Recognizing the diverse characteristics and needs of students who participate in the Smarter Balanced assessments, the Smarter Balanced member states worked together, through the Test Administration/Student Accessibility Work Group, to develop the *Accessibility and Accommodations Framework* (Smarter Balanced, 2014a; http://www.smarterbalanced.org/wp-content/uploads/2015/09/Accessibility-and-Accommodations-Framework.pdf), which guided Smarter Balanced as it worked to reach agreement on the specific universal tools, designated supports, and accommodations that would be available for the assessments. This work also incorporated research and practical lessons learned through Universal Design[24], accessibility tools, and accommodations (Thompson, Johnstone, & Thurlow, 2002).

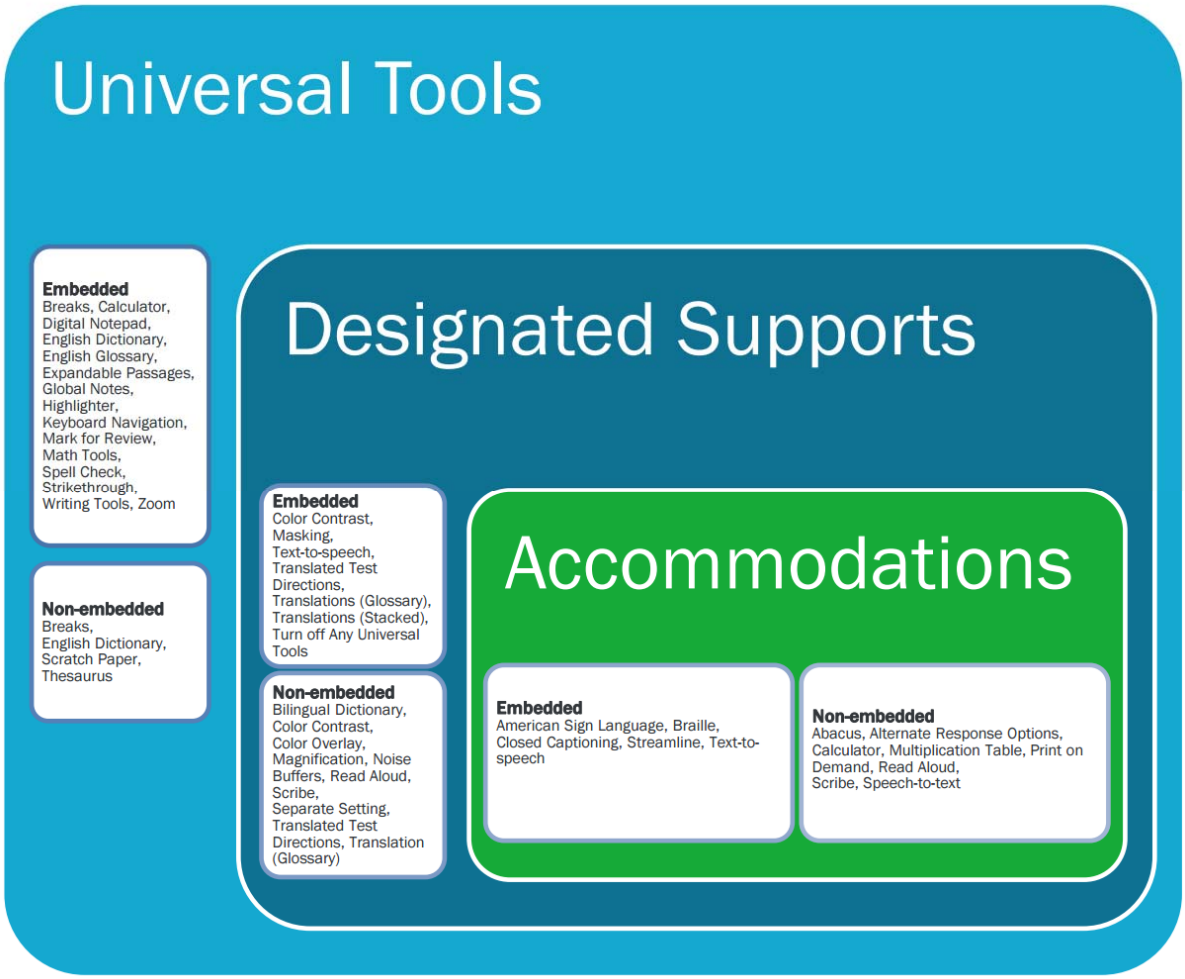
Smarter Balanced recognizes that the valid use of assessment results depends on each student having appropriate universal tools, designated supports, and accommodations when needed, based on the constructs being measured by the assessment. The Smarter Balanced assessments use technology that is intended to deliver assessments that meet the needs of individual students to help ensure that the test is fair. Online/electronic delivery of the assessments helps ensure that students are administered a test that is individualized to meet their needs while still measuring the same construct across students.

A complete list of universal tools, designated supports, and accommodations that are available for these assessments, with descriptions of each and recommendations for their use, can be found in the *Usability, Accessibility, and Accommodations Guidelines* (UAAG) (Smarter Balanced, 2017;

[24] Universal Design emphasizes the need to develop tests that are as usable as possible for all test takers in the intended test population, regardless of characteristics such as gender, age, language background, culture, socioeconomic status, or disability. For more information on Universal Design, see http://www.udlcenter.org/aboutudl and https://nceo.info/Resources/publications/OnlinePubs/Synthesis44.html.

https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf). The conceptual model underlying these guidelines is shown in Figure 2.2. Conceptual model underlying the Usability, Accessibility, and Accommodations Guidelines. This figure portrays several aspects of the Smarter Balanced assessment resources: universal tools (available for all students), designated supports (available when indicated by an adult or team), and accommodations as documented in an Individualized Education Program (IEP) or 504 plan. A universal tool or a designated support may also be an accommodation, depending on the content target and grade. This approach is consistent with the emphasis that Smarter Balanced has placed on the validity of assessment results coupled with access. Universal tools, designated supports, and accommodations are all intended to yield valid scores.

Figure 2.2. Conceptual model underlying the Usability, Accessibility, and Accommodations Guidelines.



Source: Smarter Balanced (2017), p. 4.

In order to adopt a common set of universal tools, designated supports, and accommodations based on research and on member states' best practices, Smarter Balanced worked with member states

and used a deliberative analysis strategy as described in *Accommodations for English Language Learners and Students with Disabilities: A Research-Based Decision Algorithm* (Abedi & Ewers, 2013) to determine which accessibility resources should be made available during the assessment and whether access to these resources should be moderated by an adult. As a result of this analysis and discussion with member states, some accessibility resources that states traditionally had identified as accommodations were instead embedded in tests or otherwise incorporated into the Smarter Balanced assessments as universal tools or designated supports. Other resources were not incorporated into the assessments because access to these resources was not grounded in research or was determined to interfere with the construct measured.

Meeting the Needs of Traditionally Underrepresented Populations

The policy decision to make accessibility resources available to all students based on need, rather than on eligibility status or student subgroup designation, reflects a belief among Smarter Balanced states that unnecessarily restricting access to accessibility resources threatens the validity of the assessment results and places students under undue stress and frustration. The following sections describe how the Smarter Balanced Accessibility and Accommodations Framework meets the needs of ELs and students with disabilities, including ELs with disabilities.

How the Framework Meets Needs of Students Who Are ELs. The needs of ELs are diverse and are influenced by the interactions of several factors, including their current levels of English language proficiency, their prior exposure to academic content and language in their native languages, the languages to which they are exposed outside of school, the length of time that they have participated in the U.S. education system, and the language(s) in which academic content is presented in the classroom. Given the unique backgrounds and needs of each student, the framework is designed to focus on students as individuals and to provide several accessibility resources that can be combined in a variety of ways. It embraces a variety of accessibility resources that have been designed to meet the needs of students at various stages in their English language development.

How the Framework Meets Needs of Students with Disabilities. The Accessibility and Accommodations Framework addresses the needs of students with disabilities in three ways. First, it provides for the use of digital test items that have been purposefully developed and designed to be rendered in a variety of ways to address a specific access need. By allowing the delivery of a given item to be tailored to an individual student's accommodation, the framework fulfills the intent of ensuring accessibility to allow students to demonstrate what they know and can do. Second, it allows for the test delivery system to activate accessibility resources for students based on the individual student's needs. Third, by allowing for a wide variety of digital and locally provided accommodations (including physical arrangements), it addresses a spectrum of accessibility resources to meet the needs of students with disabilities.

The Individual Student Assessment Accessibility Profile (ISAAP). Typical practice has frequently required schools and educators to document the need for specific student accommodations for an assessment, and then to document the use of those accommodations after the assessment. For example, a student's need for a large-print version of a test was indicated, and then, following the test administration, the school documented whether the student received the accommodation, whether the student actually used the large-print version, and whether any other accommodations were provided.

To facilitate the decision-making process around individual student accessibility needs, Smarter Balanced has established the Individual Student Assessment Accessibility Profile (ISAAP) tool. This

tool is designed to facilitate selection of the universal tools, designated supports, and accommodations that meet an individual student's access needs for the Smarter Balanced assessments, supported by the UAAG. If these needs are documented prior to test administration, a digital delivery system can activate the specified tools, supports, and/or accommodations when the student logs in to an assessment. By documenting the accessibility resources that are required for valid assessment of an individual student, the ISAAP allows school-level personnel to focus on individual students in a way that is efficient to manage.

The conceptual model shown in Figure 2.2 provides a structure that assists in identifying which accessibility resources should be made available for each student. In addition, the framework is designed to differentiate between universal tools that are available to all students and accessibility resources that must be assigned to individual students before the administration of an assessment.

Smarter Balanced developed the UAAG to guide the selection and administration of universal tools, designated supports, and accommodations for its assessments. The UAAG provides information, for classroom teachers, English language development educators, special education teachers, and related services personnel, on selecting and administering universal tools, designated supports, and accommodations for students who need them to use on the Smarter Balanced summative and interim assessments in ELA/literacy and mathematics. It emphasizes an individualized approach to the implementation of assessment practices for students who have diverse needs and who participate in large-scale assessments, and it supports important instructional decisions about accessibility for students. It recognizes the critical connection between accessibility in instruction and accessibility during assessment.

Fairness as a Lack of Measurement Bias: DIF Analyses

As part of the validity evidence based on internal structure, DIF analyses were conducted on Smarter Balanced items, using data from the 2014 Field Test. Again, it is important to note that the item pool for the interim assessments is derived from the same field-tested item pool used to build the summative assessments. DIF analyses are used to identify items for which identifiable groups of students (e.g., males, females) with the same underlying level of ability have different probabilities of answering an item correctly or of attaining a given score level. Ongoing study and review of findings to inform iterative, data-driven decisions is part of the Smarter Balanced framework.

Items that are found to be more difficult for some groups of students than for other groups of students may not necessarily be unfair. Fairness does not require that all groups have the same average item score. Rather, fairness requires ensuring that differences in response patterns are valid. Evaluations of validity include examination of differences in responses for groups of students who are matched on overall ability. An item would be unfair if the source of the difficulty were not a valid aspect of the item. If differences in difficulty across groups reflect real and relevant differences among those groups' levels of mastery of the tested CCSS, the item should be considered fair.

A DIF analysis asks whether, if focal-group and reference-group students of the same overall ability (as indicated by their performance on the full test) are compared, any test items are appreciably more difficult for one group, compared to other groups. DIF, in this context, is viewed as a potential source of invalidity. Table 2.1 shows the focal and reference groups used in Smarter Balanced DIF analyses.

Table 2.1. Definitions of focal and reference groups.

Group Type	Focal Groups	Reference Groups
Gender	Female	Male
Ethnicity	African American	White
	Asian/Pacific Islander	
	Native American/Alaska Native	
	Hispanic	
Special Populations	Limited English Proficient	English Proficient
	Individualized Education Program (IEP)	No IEP
	Title I	Not Title I

Analysis of the 2014-15 and 2015-16 item pools found a relatively small number of items that showed performance differences between student groups.[25] All items had previously undergone bias and sensitivity reviews. Content editors inspected flagged items, and these items were reviewed by committees of content and accessibility professionals during data review meetings, and either accepted (if the statistical differences were not caused by content issues or bias issues) or rejected (if the item was found to be flawed). Only items approved by these multidisciplinary panels of experts are eligible to be used on Smarter Balanced assessments.

Summary of Test Fairness and Implications for Ongoing Research

Many features of the Smarter Balanced assessments support equitable assessment across all groups of students. Both summative and interim assessments are developed using the principles of evidence-centered design and universal test design. Test accommodations are provided for students with disabilities, and language tools and supports have been developed for ELs. Smarter Balanced has also developed a set of guidelines to facilitate student access to the assessments guidelines for item development that aim to reduce construct-irrelevant language complexities for ELs; and comprehensive guidelines for bias and sensitivity[26]. In addition, measurement bias was investigated using DIF methods. This effort to ensure test fairness underscore the Smarter Balanced commitment to fair and equitable assessment for all students, regardless of their gender, cultural heritage, disability status, native language, or other student characteristics.

[25] For more information on these analyses, refer to pp. 3-10–3-18 of the Smarter Balanced 2014-15 and 2015-16 Technical Report Interim Assessments: https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf.

[26] Links to these guidelines are available in the Appendix.

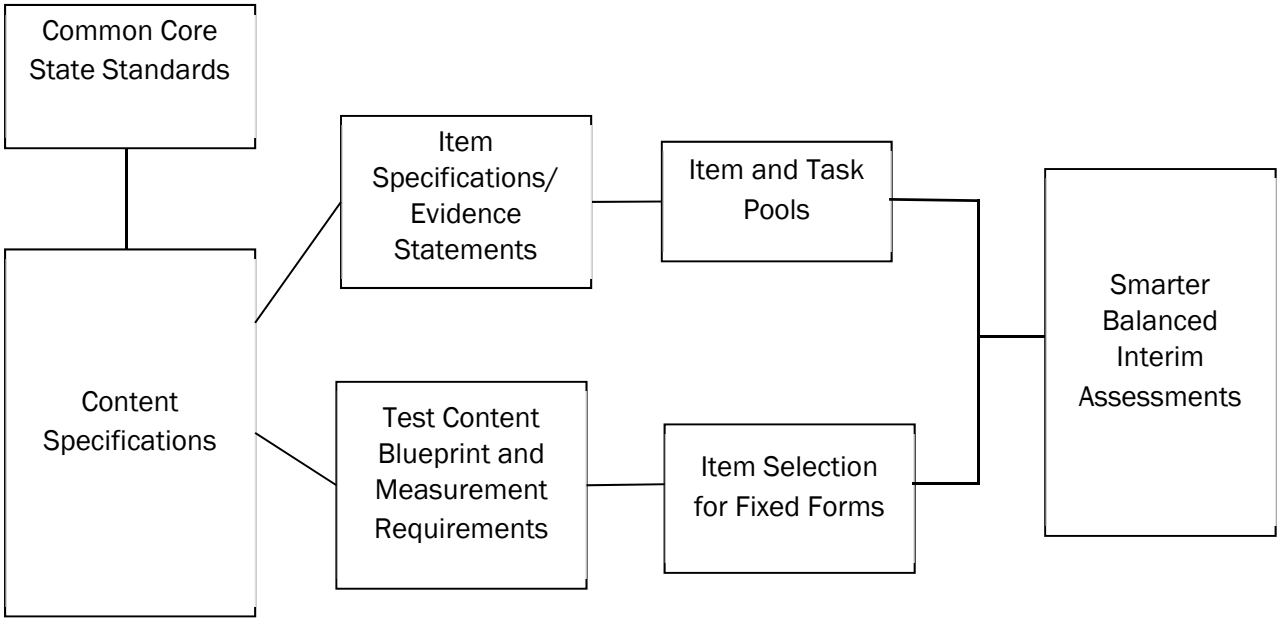
Chapter 3 – Test Design

Introduction

The primary mechanism to ensure that Smarter Balanced's measurement properties reflected the expectations of content, rigor, and performance that comprise the CCSS was to ensure alignment of the Smarter Balanced content specifications with the CCSS. Each Smarter Balanced item is aligned to a specific claim and target and to a Common Core standard. Figure 3.1 illustrates what this looks like for the interim assessments.

Figure 3. briefly encapsulates the Smarter Balanced interim assessment design process.

Figure 3.1. Components of Smarter Balanced interim assessment design.



Smarter Balanced Content Structure

Because the CCSS were not specifically developed for assessment, they contain extensive rationales for and information concerning instruction. Therefore, adopting previous practices used by many state programs, Smarter Balanced content experts distilled assessment-focused elements from the CCSS to produce content specifications for ELA/literacy (Smarter Balanced, 2015a) and mathematics (Smarter Balanced, 2015b). The content specifications for the interim assessments are the same as those for the summative assessments.

Within each of the two subject areas in grades 3–8 and high school are four broad claims. Within each claim are assessment targets. The claims in ELA/literacy and mathematics are displayed in Table 3.1.

Table 3.1. Claims for ELA/literacy and mathematics.

Claim	ELA/Literacy	Mathematics
1	Reading	Concepts and Procedures
2	Writing	Problem Solving
3	Speaking/Listening	Communicating Reasoning
4	Research	Modeling and Data Analysis

Currently, only the Listening element of ELA/Literacy Claim #3 is assessed. For the ICA Mathematics Claims #2 and #4 are reported together, which results in three reporting categories for mathematics. For the IABs, the claim/target structure is used in test specification, and results are reported for the overall IAB.

Because of the breadth of coverage of the individual claims, targets, and target clusters within each claim, evidence statements define more specific performance expectations. The relationship between targets and CCSS elements is made explicit in the Smarter Balanced content specifications. The claim/target hierarchy (see Figure 1.1 on page 5) is the basis for the structure of the summative assessments and the ICAs. IABs are based on target clusters or content domains that correspond to skill clusters commonly taught as a group.

Item Development to Content Standards

The Smarter Balanced item and task specifications for ELA/literacy and mathematics (available at http://www.smarterbalanced.org/assessments/development/) provide guidance on how to translate the content specifications for ELA/literacy (Smarter Balanced, 2015a) and mathematics (Smarter Balanced, 2015b) into actual assessment items. In addition, guidelines for bias and sensitivity (ETS, 2012); usability, accessibility, and accommodations (Smarter Balanced, 2017); and style help item developers and reviewers ensure consistency and fairness across the item bank. These specifications and guidelines have been reviewed by staff in member states, school districts, and higher education, as well as by other stakeholders. The item and task specifications describe the evidence to be elicited and provide sample task models to guide the development of items that measure student performance relative to the assessment targets.

The same blueprints are used for the summative assessments and for the ICAs. Specifically, the *ELA/Literacy Summative Assessment Blueprint* (Smarter Balanced, 2016a; https://portal.smarterbalanced.org/library/en/elaliteracy-summative-assessment-blueprint.pdf) and the *Mathematics Summative Assessment Blueprint* (Smarter Balanced, 2016d; https://portal.smarterbalanced.org/library/en/mathematics-summative-assessment-blueprint.pdf) describe the content of the ELA/literacy and mathematics ICAs, respectively, for grades 3–8 and high school—and how that content will be assessed. These blueprints also describe the composition

of both ICA components—non–performance task (non-PT) items and performance tasks (PTs)[27]—and how the results of these ICA components will be combined for score reporting. On the ICAs, PTs act in concert with non-PT items to meet the blueprints.

The *English Language Arts/Literacy Interim Assessment Blocks Fixed Form Blueprint* (Smarter Balanced, 2016b; http://www.smarterbalanced.org/wp-content/uploads/2015/09/ELA-Interim_Assessment_Blocks_Blueprint.pdf) and the *Mathematics Interim Assessment Blocks Blueprint* (Smarter Balanced, 2016c; http://www.smarterbalanced.org/wp-content/uploads/2015/09/Math_Interim_Assessment_Blocks_Blueprint.pdf) describe the content of the IABs. The IAB blueprints were created by Smarter Balanced content experts, working closely with content experts from member states. These blueprints are designed to fit typical instructional models or teaching units.

ICA Components

The ICAs for each subject consist of two parts—non-PT items and a PT—and should be administered according to the guidance provided in the *State Member Procedures Manual* (Smarter Balanced, 2015c). When administered in accordance with that guidance, ICA scores:

- Accurately describe student achievement and can be used to describe student learning in comparison to summative assessment results;
- Provide valid, reliable, and fair measures of students' progress toward, and attainment of, the knowledge and skills required to be college- and career-ready;
- Measure the breadth and depth of the CCSS across the spectrum of student ability by incorporating a variety of item types (including items and tasks scored by expert raters) that are supported by a comprehensive set of accessibility resources; and
- Utilize PT data to provide a measure of a student's ability to integrate knowledge and skills.

Test Blueprints

Both test specifications and test blueprints define the knowledge, skills, and abilities to be measured during each student's test event. A blueprint also specifies how skills are sampled from a set of content standards (i.e., for the Smarter Balanced assessments, the CCSS), as well as specifying other important factors such as DOK. A test blueprint is a formal document that guides the development and assembly of an assessment by explicating the following types of essential information:

- content (claims and assessment targets) for each assessed subject and grade,
- the relative emphasis of content standards, generally indicated as the number of items or percentage of points per claim and per assessment target;

[27] Smarter Balanced developed many different types of items beyond the traditional selected-response item, in order to measure claims and assessment targets with varying degrees of complexity by allowing students to respond in a variety of ways, rather than simply by recognizing a correct response. The different types of items and tasks that appear on the interim assessments can be viewed at http://sampleitems.smarterbalanced.org/BrowseItems. Tables showing the distribution of these item types and tasks on the interim assessments can be found on pp. 4-21–4-25 of the *Smarter Balanced Technical Report* at https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf.

- item types used or required, which communicates to item developers how to measure each claim and assessment target, and communicates to teachers and students about learning expectations; and
- Depth of Knowledge (DOK), indicating the complexity of item types for each claim and assessment target.

The ICAs are composed of non-PT and PT components. For the ICAs, responses from both components are combined to cover the test blueprint for a grade and content area and are used to produce the overall and claim scores.

Operational Interim Assessment Blueprints

ICAs. The same blueprints that were used for the summative assessments were used for the ICAs. These blueprints can be found in Appendix B of the *Smarter Balanced Technical Report* (Smarter Balanced: https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf). For each designated grade range (3–5, 6–8, and high school), the blueprint summarizes the claim score, reporting category, content category, stimuli used, items by assessment type (non-PT or PT), and total number of items by claim. Details are given separately for each grade and include claims, assessment targets, DOKs, assessment types (non-PT/PT), and the total number of items. Assessment targets are nested within claims and represent a more detailed specification of content. In addition to the nested hierarchical structure, each blueprint also specifies a number of rules applied at the global or claim levels. Most of these specifications are in the footnotes to the blueprints, which constitute important parts of the test designs.

IABs. Blueprints for all IABs are also found in Appendix B, and test maps for all assessments are found in Appendix C, of the *Smarter Balanced Technical Report:* https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf). The blueprints present the specific blocks that are available by grade level for ELA/literacy and for mathematics, beginning at grade 3 and continuing through high school. Each IAB blueprint contains information about the claim(s), assessment target(s), and DOK level(s) addressed by the items in that block, as well as the numbers of items allocated to each of those categories. Other, more content-area-specific information is also included. For example, the ELA/literacy blueprint incorporates details on passage length and scoring of responses, while the mathematics blueprint specifies to what extent the relevant task models are represented in each block. Details are given separately for each grade and include claim, assessment target, DOK, assessment type (non-PT/PT), and the total number of items.

Non–PT vs PT Test Components

No administration order is imposed on ICA components; either the non-PT portion or the PT portion of the ICAs can be administered to students first. PTs measure a student's ability to integrate knowledge and skills across multiple standards. They measure capacities—such as essay writing, research skills, and complex analysis—that are not as easy to assess with individual, discrete items.

PTs within IABs are usually stand-alone tasks with 4–6 items. They are administered separately from non-PT items. IAB PTs may be provided as practice for students to engage with the CCSS, or as professional development for teachers in hand-scoring and in understanding task demands as well as the demands of standards.

Item and Task Development

To ensure that the interim assessments measure the intended claims, the test development cycle for these assessments is iterative, involving experts from various education-related fields; is based on assessment-related research and best practices; and is identical to those for the summative assessments. For more information on the test development process, see Chapter 4 of the *Smarter Balanced 2015-16 Technical Report* (Smarter Balanced, 2016; https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf).

Item and Task Specifications

The item and task specifications for the interim assessments are the same as those for the summative assessments. These specifications help connect the content specifications and the achievement level descriptors (ALDs)[28] to the assessment itself. While the content specifications establish the claims and the types of evidence or targets that need to be collected to support these claims, more specificity was needed to develop items and tasks that appropriately and adequately measured the claims across the achievement continuum. The item and task specifications provided that specificity.

These specifications (available at http://www.smarterbalanced.org/assessments/development/) are designed to ensure that assessment items measure the assessment's claims. Indeed, the purpose of item and task specifications is to define the characteristics of items and tasks that will provide evidence to support one or more claims. To do this, the item and task specifications delineate types of evidence that should be elicited for each claim within a grade level, and provide explicit guidance on how to write items to elicit the desired evidence.

Item and task specifications provide guidelines on how to create items that are specific to each claim and assessment target. In mathematics, task models describe the knowledge, skills, and processes being measured by each of the item types aligned to particular targets, an item's or task's key features, and exemplar items. In addition, task models sometimes provide examples of plausible distracters. In ELA/literacy, the item specifications include all the necessary guidelines for item development.

Task models were developed, for each grade level and target, to delineate the expectations of knowledge and skill to be included on test questions in each grade. In addition, both ELA/literacy and mathematics item and stimulus specifications provide guidance about grade appropriateness of task and stimulus materials (the materials that a student must refer to in working on a test question). The task and stimulus models also provide information on characteristics of stimuli or activities that should be avoided because they are not germane to the knowledge, skill, or process being measured.

As previously mentioned, Smarter Balanced uses Universal Design principles to develop items that are accessible to the widest range of students possible. The underlying concept of Universal Design is to create items that accurately measure the assessment targets for all students. At the same time, Universal Design recognizes that one approach rarely works for all students. Instead, Universal

[28] Specifications of the knowledge and skills that students display at each of four different levels of achievement (Level 1, Level 2, Level 3, and Level 4). Smarter Balanced refers to these categories as numbered levels, but Smarter Balanced member states refer to them in different ways, such as "novice," "developing," "proficient," and "advanced." Students performing at Levels 3 and 4 are considered on track to demonstrating the knowledge and skills that are necessary for college and career readiness.

Design acknowledges "the need for alternatives to suit many different people" (Rose & Meyer, 2000, p. 4).

To facilitate the application of Universal Design principles, item writers are trained to consider the full range of students who may answer a test question. A simple example of this is to consider the use of vocabulary that is expected to be understood by all grade 3 students versus only those grade 3 students who play basketball. Almost all grade 3 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 such as "double dribble," "layup," "zone defense," or "full-court press."

Item specifications discuss accessibility issues that are unique to the creation of items for a particular claim and/or assessment target. These accessibility issues involve supports that various groups of students may need to access item content. By considering the supports that may be needed for each item, item writers are able to create items that can be adapted to a variety of student needs.

The use of Universal Design principles allows the Consortium to collect evidence on the widest possible range of students. By writing items that adhere to its item and task specifications, Smarter Balanced can assure member states that its assessments measure the claims and assessment targets established in the content specifications, as well as the knowledge, skills, and processes found in the CCSS, for *all* students for whom the assessment is appropriate.

Performance Task Design

A key component of college and career readiness is the ability to integrate knowledge and skills across multiple content standards. Smarter Balanced derives inferences about this ability through PTs. PTs, which are available on both the summative and interim assessments, are intended to challenge students in applying their knowledge and cognitive skills to complex, contextually rich tasks and to represent students' competence in applying that knowledge and those skills to solve substantive, meaningful problems. They give students opportunities to demonstrate their abilities to find, organize, and/or use information to solve problems, undertake research, frame and conduct investigations, analyze and synthesize data, and/or apply learning to novel situations. These activities are meant to measure capacities such as depth of understanding, writing or research skills, mathematical modeling, and complex analysis.

A Smarter Balanced PT involves student interaction with stimulus materials and/or engagement in a problem solution, ultimately leading to an exhibition of the student's application of knowledge and skills. Each PT consists of collections of questions and activities coherently connected to a single stimulus. Stimuli include a variety of forms of information (e.g., readings, video clips, data), as well as an assignment or problem situation. PTs are administered online via computer (they are not computer adaptive) and require one to two class periods to complete. PTs are an integral part of the Smarter Balanced test design.

PTs are constructed so that they can be delivered effectively in a school or classroom environment (Dana & Tippins, 1993). They adhere to specifications used by item writers to develop new tasks that focus on different content but are comparable in contribution to the blueprint. Task specifications include, but are not limited to, pre-assessment classroom activities, materials and technology needs, and allotted time for assessment. IABs that are based only on PTs yield indicators of Below Standard, At or Near Standard, or Above Standard.

Item/Task Pool Specifications

An *item pool* refers to a collection of test questions (known as items) that supports the test blueprint for a particular content area and grade. Smarter Balanced has taken multiple steps to ensure the quality of the items in its item pool. Building on the ongoing process of developing item/task specifications and test blueprints, Smarter Balanced uses an iterative process for creating and revising each item as well as for the collection of items. Items were tested and refined using three steps: small-scale tryouts, a large pilot test, and a large field test. Details of the pilot and field tests done for the 2013-14 assessments, for example, can be found in the *2013-14 Technical Report* (Smarter Balanced, 2016e; https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf). During each phase of the pilot and field tests, cognitive laboratories were used to understand the strategies that students used to respond to the items. By incorporating this tiered and iterative approach, the item and task specifications that guided the development of the final operational pool, from which both summative and interim assessment items are drawn, were improved based on lessons learned during tryouts.

Using test blueprints, measurement experts specified the numbers and distributions of items to be written for each content area and grade. Pools of items and tasks were written specifically to support the proportions of items and the intended difficulty distributions in the operational blueprints. Teachers were integrally involved throughout the creation of the item/task pool. Some teachers participated in the processes described in the flow charts that appear in Appendix D of the *2014-15, 2015-16 Interim Assessment Technical Report* (Smarter Balanced) https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf). Others developed items, through a rigorous item-writing process, or reviewed the items for accuracy and appropriateness of the content knowledge and skill levels required for students to respond to the items. Teams of content experts reviewed items for potential issues of bias in favor of or against any demographic group of students, and for accessibility for students with disabilities and ELLs. Content, bias, and accessibility reviews were conducted prior to items' administration to any students. Following pilot and field test administrations, items were again reviewed if administration data indicated a potential problem with the item. Finally, teachers participated in range-finding and scoring of constructed-response items[29] and tasks, to ensure that the items/tasks could be properly scored given their scoring rubrics.

Content Alignment for ICAs

Smarter Balanced is committed to ensuring that its measurement reflects the CCSS expectations of content, rigor, and performance. To that end, Smarter Balanced has designed item specifications to demonstrate alignment through methodologies that reflect evidence-centered design theory. Test content alignment is at the core of content validity and consequential validity[30] (Martone & Sireci, 2009).

Webb (1997) identified several categories of criteria for judging content alignment. The Smarter Balanced alignment study conducted by HumRRO (HumRRO, 2016) describes how well the Smarter Balanced summative and ICA test designs address expectations embodied in the content

[29] Items, tasks, or exercises for which test takers must create their own responses or products rather than choose a response from a specified set. Answer rubrics are developed during the item development process for these types of items.
[30] Consequential validity describes the aftereffects and possible social and societal results from a particular assessment or measure.

specifications and the CCSS. The HumRRO study directed attention to test alignment in addition to individual item 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 pool of possible items and tasks. For inferences from test results to be justifiable, that sampling of items must be an adequate representation of the broad domain, providing strong evidence to support claims based on the test results.

Assessment is constrained to some extent by time and resources. Items and tasks that require extensive time (performance tasks and constructed responses), items that require extensive scoring, and items that require a lot of computer bandwidth (videos, animations) must be used on a limited basis and must be chosen carefully. Smarter Balanced content experts have scrutinized each blueprint to assure optimal content coverage and prudent use of time and resources, and the blueprints were subject to member-state approval through a formal vote. In general, the Smarter Balanced summative and ICA blueprints represent content sampling proportions that reflect intended emphasis in instruction and mastery at each grade level. Specifications for numbers of items by claim, assessment target, DOK, and item type demonstrate the desired proportions within test delivery constraints.

IAB Design

IABs are designed by teams of content experts to reflect groups of related skills that are most likely to be addressed in instructional units. These tests are short and contain a focused set of skills. They indicate whether a student is clearly above or below standard.

Summary of chapter on Test Design

The intent of this chapter is to show how the design of the Smarter Balanced interim assessments supports the purposes of these assessments. Content specifications have been derived directly from the CCSS, expressing the standards as measurable elements made explicit in the structure of Smarter Balanced claims and assessment targets. Building on these specifications, test blueprints provide appropriate proportions of CCSS content coverage. Using the blueprints, item writers write items and tasks in quantities that support non-PT and PT delivery. Expansion of item and task types has facilitated student responses that provide more insight into proficiency than those provided by selected-response items alone. The use of PTs in ICAs and in some IABs addresses the need to assess application and integration of skills.

Chapter 4 – Test Administration

Introduction

According to the *Standards*, "The usefulness and interpretability of test scores require that a test be administered and scored according to the developer's instructions" (AERA et al., 2014, p. 111). Smarter Balanced created and disseminated a customizable test administration manual, (TAM) to ensure standardized test administration procedures and, thus, uniform test administration conditions for all students in Smarter Balanced member states. This chapter describes the TAM and Smarter Balanced test administration procedures and conditions.

For fixed-form tests (the Smarter Balanced interim assessments are currently only available as fixed forms), information on reliability and error is derived from the properties of items on the test forms. This information is valid only when tests are administered in the standardized manner described in the TAM and only for the first administration of the tests. In many cases, these tests are administered in a non-standard manner: they may be administered at multiple points in the school year, and items on the IABs and ICAs may be used in other contexts. Additionally, some teachers use the tests as a tool for discussion, working through each item with the class. In these cases, the items have been exposed and no longer have the same parameters as before they were first administered. When tests are used in these manners, results can still be useful if test-taking conditions are known. When interim scores are recorded for future reference, the conditions of administration should be noted, so that results can be interpreted appropriately and accurately.

Test Administration Specifications

The ICAs and the IABs may be administered online through the Smarter Balanced open-source test administration system or through an alternative service provider. The ICAs use the same specifications as the summative assessments. The specifications for the IABs have the following additional content allowances:

- Multiple configurations: Districts and schools may elect to administer both ICAs and IABs during the school year.
- Multiple administrations: ICAs and IABs (including those for the same block of skills) may be administered multiple times within an academic year.
- Smarter Balanced member states may determine the schedule for interim assessment administration, or may delegate the administration of interim assessments to schools/districts.
- Smarter Balanced does not limit the number of times that ICAs and/or IABs are administered.

Given these allowances, members should be aware that testing multiple times a year limits the item pool available to students, which will increase the possibility of students encountering the same item several times Overexposed items are unlikely to hold their original parameters and may skew performance results. To prevent this, schools and classrooms may want to limit their testing program to either judicious use of ICAs or coordinated use of IABs.

Administration Instructions

The *State Member Procedures Manual* (Smarter Balanced, 2015c) provides a high-level overview of the assessment system (including summative and interim assessments), with expected policies and procedures for administration, required trainings, general information about the open-source administration platform, information about the evidence that states must provide to Smarter Balanced annually, procurement information, and links to resource documents.

Specific components of the *Procedures Manual* require customization to meet the unique needs of each member state. Member states have developed specific instructions for administration of Smarter Balanced interim assessments, customized to fit the states' needs, and have made these instructions available on their websites.

Clear Directions to Ensure Uniform Administration

Smarter Balanced Test Administration Manuals (TAMs), developed by and specific to each member state, (Smarter Balanced, Assessment Consortium 2014) include instructions that clearly articulate various aspects of the administration process. The TAMs, cover an extensive amount of material related to events that occur before, during, and after testing. In addition, the TAMs point users to training materials that provide further detail and clarity to support reliable test administration by qualified test administrators. The TAMs describe the general rules of online testing, including pause rules; scheduling tests; recommended order of test administration; assessment duration, timing, and sequencing information; and the materials that the test administrator and students need for testing.

Responsibilities of Test Administrators

The Standards (AERA et al., 2014) provide guidance to test administrators and test users. Test administrators are guided to carefully follow the standardized procedures (Standard 6.1); inform test takers of available accommodations (Standard 6.2); report changes or disruptions to the standardized test administration (Standard 6.3); furnish a comfortable environment with minimal distractions (Standard 6.4); provide appropriate instructions, practice, and other supports (Standard 6.5); and ensure the integrity of the test by eliminating opportunities for test-taker malfeasance (Standard 6.6). In addition, test users are responsible for test security at all times (Standard 6.7). However, interim tests are not secure in the same manner as summative tests. Although they are secure in the sense that they are not to be made public, through social media or other means, the items on interim assessments may be discussed or used as classroom examples.

Chapter 5 – Reporting and Interpretation

Introduction

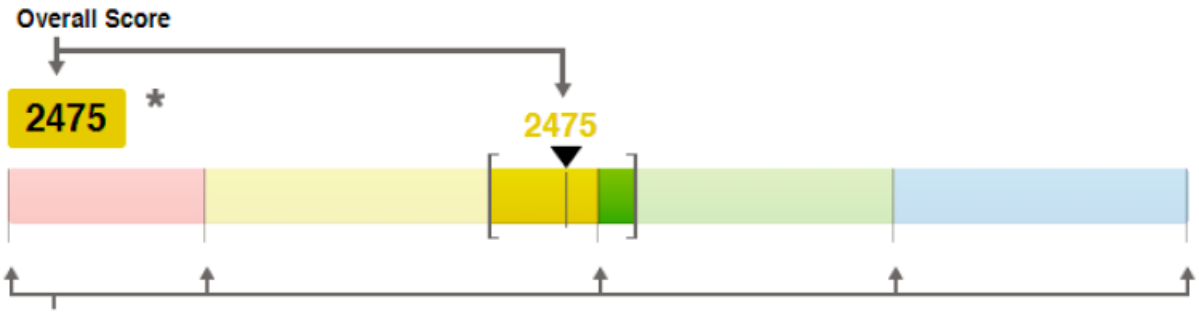
Member states' use of the Smarter Balanced reporting system is optional and configurable; therefore, information about a specific member state's reports should be gathered from the state's websites and materials. Scores from the ICAs provide information about student achievement regarding college and career readiness. As noted in Chapter 3, the ICAs provide an overall indicator of proficiency and a set of claim indicators corresponding to broad areas within the content areas. Information from IABs is reported in the same manner as claim-level information on the ICAs. The Consortium provides a set of reports based on these scores and on claim information that members may customize for their own use. This chapter provides an overview of the Smarter Balanced reporting system. For detailed information on this system, consult the *Reporting System User Guide* (Smarter Balanced, 2014z).

Overall Test Scores

Scale scores are the basic units of overall reporting for the Smarter Balanced summative assessments and ICAs. These scores fall along a continuous vertical scale (from approximately 2000 to 3000) that increases across grade levels. Scores are used to describe an individual student's current level of achievement. They can also be used to track growth over time. When aggregated, scale scores are used to describe achievement for different groups of students.

For the ICAs, the Smarter Balanced reporting system communicates an overall scale score in relation to defined achievement levels, using graphics similar to Figure 5.. By default, the system uses generic terms for the achievement levels—Level 1, Level 2, Level 3, and Level 4—but members may customize their reporting levels by using terms such as "novice," "developing," "proficient," and advanced," or others.

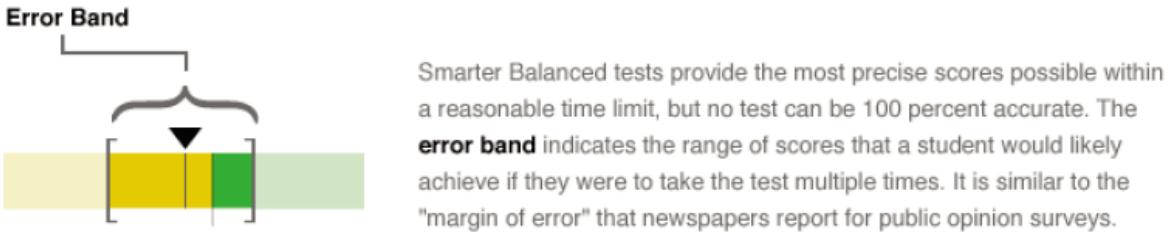
Figure 5.1. Portrayal of ICA score reporting levels.



Source: Smarter Balanced (2014z), p. 13.

In Figure 5.1, the overall score is 2475, which is in Level 2, and the score's error band encompasses Level 3. Smarter Balanced reporting provides information to help users understand the meaning of the error bands, as shown in Figure 5..

Figure 5.2. EXPLANATION OF ERROR BANDS DISPLAYED ON SMARTER BALANCED REPORTS.



Source: Reporting System User Guide, p.120.

Depicting errors and error bands in score reporting is an important measurement principle. In Figure 5.1 and Figure 5.2, the score is represented by the vertical line and black triangle. The error band is shown by the brackets. If the test were to be given again, the subsequent score would be likely to fall within this band.

Smarter Balanced has developed a set of optional reporting ALDs, for ELA/literacy and mathematics, that are aligned with the CCSS and with Smarter Balanced assessment claims. The intent of these ALDs is to specify, in content terms, the knowledge and skills that students may display at four levels of achievement. The full set of optional Reporting ALDs is shown in Appendix E of the *2014-15 and 2015-16 Technical Report* (Smarter Balanced) https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf).

Sub-Scores

ICA claim and reporting category indicators are based on sub-scores for important domain areas within each content area. The claims and reporting categories are primary structural elements in ICA test blueprints. IAB results are based on the IAB blueprints, which reflect instructional targets or domain groupings.

Table 5. and

Table 5. provide the claims or sub-score reporting categories for ICAs.

Table 5.1. ELA/literacy claims.

Claim #1—Reading • Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
Claim #2—Writing • Students can produce effective and well-grounded writing for a range of purposes and audiences.
Claim #3—Speaking/Listening • Students can employ effective speaking and listening skills for a range of purposes and audiences. *Currently, only listening is assessed.*
Claim #4—Research • Students can engage in research/inquiry to investigate topics and to analyze, integrate, and present information.

Table 5.2. Mathematics claims and score reporting categories.

Claim #1—Concepts and Procedures • Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
Claim #2—Problem Solving/Claim #4—Modeling and Data Analysis • Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies. • Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems. • Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.
Claim #3—Communicating Reasoning • Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.

These designations are based on the distance of the score from the Level 2/3 cut score[31] in terms of the standard error of the score. An error band of plus and minus 1.5 standard errors is constructed around the score. The score is designated "Above," "Near," or "Below," if the Level 2/3 cut is, respectively, below, within, or above the error band.

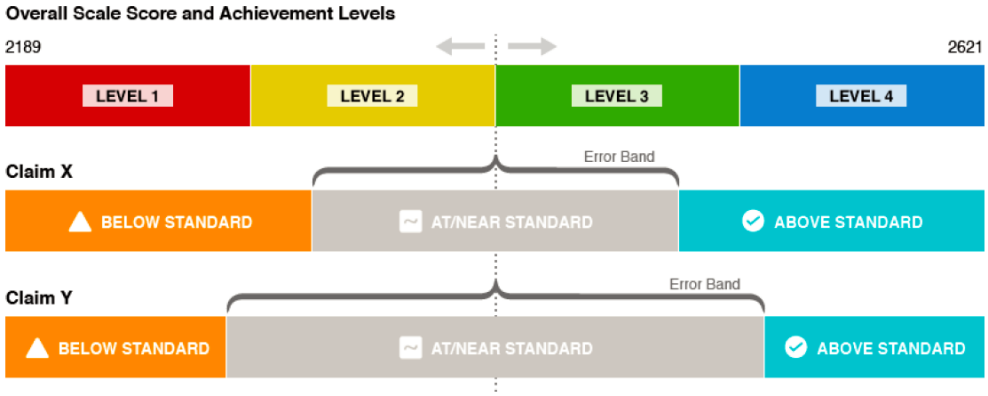
[31] Cut scores are selected points on the score scale of a test based on a generally accepted methodology and reflect the judgments of qualified people.

Table 5.3. IAB score and ICA sub-score categories.

Above Standard	Score is > 1.5 SEMs or more above the Level 2/3 cut score
At or Near Standard	The Level 2/3 cut score falls within an error band of +/- 1.5 SEMs around the sub-score
Below Standard	Score is > 1.5 SEMs or more below the Level 2/3 cut score

Figure 5.3 displays these distinctions visually. Instead of using error bands, it shows the reporting level area that would result from a scale score and an SEM.

Figure 5.3. Portrayal of IAB score and ICA sub-score reporting.



Source: Smarter Balanced (2014z), pp. 116–117.

IAB scores and ICA sub-scores are portrayed in Smarter Balanced reports using the three-level structure shown in Figure 5.3 (also called "traffic-light" indicators); sub-score scale scores and SEMs are available to Smarter Balanced member states in the data provided by their assessment vendors.

Results for IABs are reported in the same manner as claim results on summative tests and ICAs. That is, the results are given as "Below Standard," "At or Near Standard," or "Above Standard." The rules for assigning a score to these categories are based on the distance of the score from the Level 2/3 cut in terms of 1.5 times the score's standard error as described above.

Reporting for IABs is focused on individual student reporting and on communicating block-level results for a list of students by assessment grade. Since IABs are locally scored, educators will see individual student responses to hand-scored items through the Scoring component.

Conclusion

Educators, students, parents, and guardians can use interim assessment results to understand a student's achievement, progress toward mastery of the CCSS, and attainment of the academic knowledge and skills required for the student to be college- and career-ready. These results may also

provide context for a parent-teacher conference, or, when used with other instructional data, may help to identify areas for instructional focus.

Appendix

References

Abedi, J., & Ewers, N (2013). Accommodations for English language learners and students with disabilities: A research-based decision algorithm. Available from http://www.smarterbalanced.org/wp-content/uploads/2015/08/Accomodations-for-underrepresented-students.pdf American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Authors.

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

"Connecting Summative Assessment to Improving Teaching and Learning" – Post-Test Workshops - Presented by ETS, CDE, and WestEd – May and June 2016.

Conley, D. T., Drummond, K. V., de Gonzalez, A., Rooseboom, J., & Stout. O. (2011). Reaching the goal: The applicability and importance of the Common Core State Standards to college and career readiness. Eugene, OR: Educational Policy Improvement Center. Cook, H.G. & McDonald, R. (2013). Tool to Evaluate Language Complexity of Test Items. Wisconsin Center for Education Research. Retrieved from https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2013_05.pdf)

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education. Dana, T. M., & Tippins, D. J. (1993). Considering alternative assessment for middle level learners. Middle School Journal, 25, 3-5.

DeMauro, G. E. (2004). Test alignment considerations for the meaning of testing. Paper presented at the CCSSO Annual Conference on Large Scale Assessment, Boston, MA. Educational Testing Service (2002). ETS standards for quality and fairness. Princeton, NJ: Author.

Educational Testing Service (ETS). (2012). Specifications for an interim system of assessment. Princeton, NJ: Author.

ETS. (2012). *Smarter Balanced Assessment Consortium: Bias and sensitivity guidelines*. Retrieved from https://portal.smarterbalanced.org/library/en/v1.0/bias-and-sensitivity-guidelines.pdf

Human Resources Research Organization (HumRRO). *Smarter Balanced Assessment Consortium: Alignment study report*. Retrieved from https://portal.smarterbalanced.org/library/en/smarter-balanced-assessment-consortium-alignment-study-report.pdf

Kane, M. (2006). Validation. In R. L. Brennan (Ed). Educational measurement (4th ed., pp. 17-64). Washington, DC: American Council on Education/Praeger. Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. Review of Educational Research, 79, 1-76. Rose, D., & Meyer, A. (2000). Universal design for learning, associate editor column. Journal of Special Education Technology 15 (1): 1-12Smarter

Balanced Assessment Consortium (Smarter Balanced). (2012). *Smarter Balanced Assessment Consortium: General accessibility guidelines*. Retrieved from https://portal.smarterbalanced.org/library/en/general-accessibility-guidelines.pdf

Smarter Balanced. (2014a). *Smarter Balanced Assessment Consortium: Accessibility and accommodations framework*. Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/09/Accessibility-and-Accommodations-Framework.pdf

Smarter Balanced Assessment Consortium: Guidelines for Accessibility for English Language Learners. Retrieved from https://portal.smarterbalanced.org/library/en/v1.0/guidelines-for-accessibility-for-english-language-learners.pdf

Smarter Balanced. (2015a). *Smarter Balanced content specifications for the summative assessment of the CCSS for English language arts/literacy*. Retrieved from https://portal.smarterbalanced.org/library/en/english-language-artsliteracy-content-specifications.pdf

Smarter Balanced. (2015b). *Smarter Balanced content specifications for the summative assessment of the CCSS for mathematics*. Retrieved from https://portal.smarterbalanced.org/library/en/mathematics-content-specifications.pdf

Smarter Balanced. (2016a). *ELA/literacy summative assessment blueprint*. Retrieved from https://portal.smarterbalanced.org/library/en/elaliteracy-summative-assessment-blueprint.pdf

Smarter Balanced. (2016b). *English language arts/literacy interim assessment blocks fixed form blueprint.* Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/09/ELA-Interim_Assessment_Blocks_Blueprint.pdf

Smarter Balanced. (2016c). *Mathematics interim assessment blocks 2016-17 blueprint*. Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/09/Math_Interim_Assessment_Blocks_Blueprint.pdf

Smarter Balanced. (2016d). *Mathematics summative assessment blueprint*. Retrieved from https://portal.smarterbalanced.org/library/en/mathematics-summative-assessment-blueprint.pdf.

Smarter Balanced*. (2016). Reporting System User Guide*. Retrieved from *http://www.smarterapp.org/manuals/Reporting-UserGuide.html*

Smarter Balanced. (2016e). *Smarter Balanced Assessment Consortium: 2013-14 technical report*. Retrieved from https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

Smarter Balanced. (2016f). *Smarter Balanced Assessment Consortium: 2014-15 technical report.* Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf

Smarter Balanced. (2017). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines*. Retrieved from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Thurlow, M. L., Quenemoen, R. F., & Lazarus, S. S. (2011). Meeting the Needs of Special Education Students: Recommendations for the Race to the Top Consortia and States. Paper prepared for Arabella Advisors Webb, N. L. (1997a, April). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph No. 6. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (March 28, 2002) Depth-of-Knowledge Levels for Four Content Areas, unpublished paper. Young, J., Pitoniak, M. J., King, T. C., & Ayad, E. (2012).

List of Acronyms

ALDs: achievement level descriptors

CAT: computer adaptive testing

CCSS: Common Core State Standards

DIF: differential item functioning

DOK: Depth of Knowledge

ELA/literacy: English language arts/literacy

ELs: English earners

IABs: Interim Assessment Blocks

ICAs: Interim Comprehensive Assessments

IEP: Individualized Education Program

ISAAP: Individual Student Assessment Accessibility Profile Non-PT: non–performance task

PT: performance task

SEM: standard error of measurement

TAM: Test Administration Manual

UAAG: Usability, Accessibility, and Accommodations Guidelines

Glossary

Accessibility supports: Universal tools, designated supports, and accommodations.

Achievement categories (also called *claim levels*): Below Standard, Near Standard, and Above Standard.

Achievement level descriptors (ALDs): Specifications of the knowledge and skills that students display at each of four levels (i.e., Level 1, Level 2, Level 3, and Level 4). Smarter Balanced refers to these categories as numbered levels, but Smarter Balanced member states refer to them in different

ways, such as “novice,” “developing,” “proficient,” and “advanced.” Students performing at Levels 3 and 4 are considered on track to demonstrating the knowledge and skills that are necessary for college and career readiness. Process for determining these levels found here: http://www.smarterbalanced.org/assessments/scores/

Aggregate score: A total score formed by combining scores on the same test or across test components

Alignment: The correspondence between student learning standards and test content.

Bias: “[C]onstruct underrepresentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers and consequently the reliability/precision and validity of interpretations and uses of their test scores.” (AERA et al., 2014, p. 216)

Claim: The concept or characteristic that an assessment is designed to measure. Also referred to as a construct.

Computer adaptive testing (CAT): A method of testing that actively adapts to the test taker’s ability level during a computerized assessment. *Consequential validity*: Consequential validity describes the aftereffects and possible social and societal results from a particular assessment or measure. For an assessment to have consequential validity, it must not have negative social consequences that seem abnormal. If this occurs, it signifies that the test is not valid and is not measuring things accurately.

Construct: The concept or characteristic that a test is designed to measure. (used interchangeably with claim in report?)

Construct-irrelevant barriers: Extraneous, uncontrolled variables that affect assessment outcomes.

Construct-irrelevant variance: The introduction of extraneous, uncontrolled variables that affect assessment outcomes. When this happens, the meaningfulness and accuracy of examination results is adversely affected, the legitimacy of decisions made upon exam results is affected, and the validity is reduced.

Content cluster: Related content or a grouping of related standards that can be measured with similar skills.

Cut score: Cut scores are selected points on the score scale of a test. The points are used to determine whether a particular test score is sufficient for a specific purpose. For example, student performance on a test may be classified into one of several categories, such as basic, proficient, or advanced, based on cut scores. The setting of cut scores on widely used tests in educational contexts requires the involvement of multiple stakeholders in a multistage, judgmental process. Cut scores should be based on a generally accepted methodology and should reflect the judgments of qualified people.

Depth of Knowledge (DOK): A ranking of tasks (from 1–4) according to the complexity of thinking required to successfully complete them

Differential item functioning (DIF): For a particular item in a test, a statistical indicator of the extent to which different groups of test takers who are at the same ability level have different frequencies of correct responses, or, in some cases, different rates of choosing various item options.

Digital Library: an online collection of instructional and professional learning resources contributed by educators for educators.

Domain: A large group of related standards.

Evidence-centered design: An approach to constructing educational assessments in terms of evidentiary arguments (See Figure 2.1 for a visual representation of this approach)
Field testing: A test administration that is used to check the adequacy of testing procedures and the statistical characteristics of new test items or new test forms. A field test is generally more extensive than a pilot test.

Fixed form: All students receive the same future questions, regardless of individual students' performance on past questions. Opposite of adaptive testing.

Formative assessment: A range of formal and informal assessment procedures (including diagnostics) conducted by teachers during the learning process to modify teaching and learning activities to improve student attainment.

Item specifications documents: Like content specifications documents, these documents provide educators with information about claims, targets, standard identification, and DOK level(s). They also list the types of evidence needed to indicate that the student has the knowledge, skills, and abilities measured by that type of question.

Item/task pool: A collection of test questions (known as items) that supports the test blueprint for a particular content area and grade.

Operational use: The actual use of a test to inform an interpretation, decision, or action, based in part or wholly on test scores.

Performance task: An assessment component that involves significant student interaction with stimulus materials and/or engagement in a problem solution, ultimately leading to an exhibition of the student's application of knowledge and skills.

Pilot test: A test administered to a sample of test takers to try out some aspects of the test or the test items, such as the instructions, time limits, item response formats, or item response options.

Range-finding: Review of responses using rubrics, to validate the rubrics and select anchor papers (models) for use during operational scoring.

Scale score: The number score.

Sensitivity: Awareness of the need to avoid explicit bias.

Standard error of measurement (SEM): The *Standards* (AERA et al., 2016) define the SEM as the standard deviation of an individual's observed scores from repeated administrations of a test (or parallel form of a test) under identical conditions. Because such data generally cannot be collected, the SEM is usually estimated from group data.

Stimuli: ELA/literacy passages.

Summative assessment: Summative assessments are used to evaluate student learning, skill acquisition, and academic achievement at the conclusion of a defined instructional period—typically at the end of a project, unit, course, semester, program, or school year.

Targets: Targets are the bridge between the content standards and the assessment evidence that supports the claim. They ensure sufficiency of evidence to justify each claim.

Test scaling: The process of creating a scale or a scale score to enhance test score interpretation, by placing scores from different tests or test forms on a common scale or by producing scale scores designed to support score interpretations.

Universal Design: Universal Design emphasizes the need to develop tests that are as usable as possible for all test takers in the intended test population, regardless of characteristics such as gender, age, language background, culture, socioeconomic status, or disability *validation*: The process of gathering evidence to support each proposed score interpretation or use.

Validity: The degree to which each interpretation or use of a test score is supported by the accumulated evidence.

Validity argument: "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system." (AERA et al., 2014, pp. 21–22).

Smarter Balanced Resources: What and Where

Bias and Sensitivity Guidelines:
(http://www.smarterbalanced.org/wp-content/uploads/2015/08/BiasandSensitivityGuidelines.pdf

2013-14 Technical Report:
https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf

Development and Design homepage (including links to test blueprints and content and item/task specifications): http://www.smarterbalanced.org/assessments/development/

ELA/Literacy Content Specifications:
https://portal.smarterbalanced.org/library/en/english-language-artsliteracy-content-specifications.pdf

Appendix B: Grade-Level Tables (from the ELA/literacy content specifications):
https://portal.smarterbalanced.org/library/en/appendix-b-grade-level-tables.pdf

Mathematics Content Specifications:
https://portal.smarterbalanced.org/library/en/mathematics-content-specifications.pdf

Interim Assessments Overview: https://portal.smarterbalanced.org/library/en/interim-assessments-overview.pdf

2016–17 Interim Assessment Blocks Overview: https://www.smarterbalanced.org/wp-content/uploads/2015/08/Interim_Assessment_Blocks_overview.pdf

ELA/Literacy Summative Assessment Blueprint:
https://portal.smarterbalanced.org/library/en/elaliteracy-summative-assessment-blueprint.pdf

Mathematics Summative Assessment Blueprint:
https://portal.smarterbalanced.org/library/en/mathematics-summative-assessment-blueprint.pdf

ELA/Literacy Interim Assessment Blocks Fixed Form Blueprint: http://www.smarterbalanced.org/wp-content/uploads/2015/09/ELA-Interim_Assessment_Blocks_Blueprint.pdf

Mathematics Interim Assessment Blocks Blueprint: http://www.smarterbalanced.org/wp-content/uploads/2015/09/Math_Interim_Assessment_Blocks_Blueprint.pdf

Reporting Scores homepage: http://www.smarterbalanced.org/assessments/scores/

Accessibility & Accommodations factsheet: http://www.smarterbalanced.org/wp-content/uploads/2015/08/SmarterBalanced_Accessibility_Factsheet.pdf

Accessibility and Accommodations Framework: http://www.smarterbalanced.org/wp-content/uploads/2015/09/Accessibility-and-Accommodations-Framework.pdf

Usability, Accessibility, and Accommodations Guidelines (UAAG): https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Alignment Study Report (includes results of the HumRRO alignment study): https://www.smarterbalanced.org/wp-content/uploads/2016/05/Alignment-Study-Report_HumRRO.pdf