

Chapter 2: Validity.....	3
Introduction	3
Essential Validity Elements for Summative and Interim Assessments	4
Table 1. Synopsis of Essential Validity Evidence Derived from <i>Standards</i> (AERA et al., 1999, p. 17).....	5
Careful Test Construction.....	5
Adequate Measurement Precision (Reliability).....	6
Appropriate Test Administration.....	7
Appropriate Scoring.....	7
Accurate Scaling and Linking.....	7
Appropriate Standard Setting.....	8
Attention to Fairness, Equitable Participation, and Access.....	9
Validating “On-Track/Readiness”	9
Adequate Test Security.....	11
Summary of Essential Validity Evidence based on the Smarter Pilot- and Field Tests.....	11
Table 2. Essential Validity Evidence for the Summative and Interim Assessments for Careful Test Construction.....	12
Table 3. Essential Validity Evidence for the Summative and Interim Assessments for Adequate Measurement Precision (Reliability).....	13
Table 4. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Test Administration.....	14
Table 5. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Scoring.....	14
Table 6. Essential Validity Evidence for the Summative and Interim Assessments for Accurate Scaling and Linking.....	15
Table 7. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Standard Setting.....	16
Table 8. Essential Validity Evidence for the Summative and Interim Assessments for Attention to Fairness, Equitable Participation and Access.....	17
Table 9. Essential Validity Evidence for the Summative and Interim Assessments for Validating “On-Track/Readiness”.....	17
Table 10. Essential Validity Evidence for the Summative and Interim Assessments for Adequate Test Security.....	18
The <i>Standards’</i> Five Primary Sources of Validity Evidence	18
Purposes of the Smarter Balanced System for Summative, Interim, and Formative Assessments	20
Table 11. Validity Framework for Smarter Balanced Summative Assessments.....	22

Table 12. Validity Framework for Smarter Balanced Interim Assessments.....	23
Evidence Using the Five Primary Sources of Validity Framework	23
Table 13. Listing of Evidence Type, Evidence Source, and Primary Validity Source for Summative, Interim, and Formative Test Purposes.....	24
Conclusion for Field Test Validity Results.....	30
References	31
American Institutes for Research (2014b). Smarter Balanced Scoring Specification: 2014–2015 Administration.....	31

Chapter 2: Validity

Introduction

Validity refers to the degree to which each interpretation or use of a test score is supported by the accumulated evidence (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999; 2014; ETS, 2002). It constitutes 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 accumulating evidence to support each proposed score interpretation or use. This validation process does not rely on a single study or gathering one type of evidence. Rather, validation involves multiple investigations and different kinds of supporting evidence (AERA, APA, & NCME, 1999; 2014; Cronbach, 1971; ETS, 2002; Kane, 2006). It begins with test design and is implicit throughout the entire assessment process, which includes item development and field-testing, analyses of items, test scaling, and linking, scoring, and reporting. This chapter provides an evaluative framework for the validation of the Smarter Balanced Assessment System. It points the reader to supporting evidence in other parts of this technical report and other sources that seek to demonstrate that the Smarter Balanced Assessment System adheres to guidelines for fair and high quality assessment. Since many aspects of the program were still under development at the time of this report, additional research that further supports the Smarter Balanced goals is mentioned as appropriate throughout this chapter.

This chapter is organized primarily around the principles prescribed by AERA, APA, and NCME's *Standards for Educational and Psychological Testing* (1999; 2014) and the Smarter *Balanced Assessment Consortium: Comprehensive Research Agenda* (Sireci, 2012), both of which serve the primary sources for this chapter. The *Standards* are considered to be "the most authoritative statement of professional consensus regarding the development and evaluation of educational and psychological tests" (Linn, 2006, p. 27) currently available. As this report and the associated validation was nearing completion, the 2014 *Standards* were published. The basic notions of validity described in the 1999 *Standards* are consistent with those entailed in the 2014 *Standards*. The 2014 *Standards* differ from earlier ones in the emphasis given to the increased prominence of technology in testing, such as computer adaptive testing (CAT) and automated scoring. CAT methodology and automated scoring approaches are both important components of the Smarter Balanced assessments. The use of the *Standards* in this chapter refers to the 2014 version unless the 1999 edition is specifically referenced.

The validity evidence presented in this technical report was collected in the context of two phases that consisted of a pilot test and field test prior to any operational administration. As a result, many critical elements of the program were being developed simultaneously within a short time span late in 2014. The validity evidence is intended to provide the best possible information for both understanding the degree to which the Smarter Balanced Consortium is meeting its goals consistent with completion of the Field Test phase, as well as the steps needed to be undertaken to improve the system as it evolves operationally.

Two types of overlapping validity frameworks are presented. The first validity framework corresponds to the essential validity elements (AERA et al. 1999, p.17). This essential validity information is more consistent with the types of evidence typically reported for many large-scale educational assessment programs. These essential validity elements present a more traditional synopsis of validity evidence, which form the basis for the evidence demonstrated for the Smarter Balanced Field Test to date and the initial operational administrations. The second more comprehensive validity framework cross-references Smarter Balanced test purposes against the *Standards'* five primary sources of validity evidence. These five sources of validity evidence consist of (1) test content, (2) response processes,

(3) internal structure, (4) relations to other variables, and (5) consequences of testing. Evidence in support of the five sources of validity will need to be addressed more fully in the course of ongoing Smarter Balanced research. The essential validity elements form a subset and sample the five sources of validity evidence. The essential validity framework is presented first followed by the five primary sources of validity.

Essential Validity Elements for Summative and Interim Assessments

The *Standards* describe the process of validation that consists of developing a sufficiently convincing argument, based on empirical evidence, that the interpretations and actions based on test scores are sound. Kane (1992, 2006) characterized this process as a validity argument, which is consistent with the validation process described by the 1999 *Standards*.

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 . . . relies on all the available evidence relevant to the technical quality of a testing system (AERA et al., 1999, p. 17).

The 1999 *Standards* describe these essential validity elements as “evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees.” Some modifications were made to the original 1999 specification of these essential elements. Careful attention to fairness for all examinees was changed to attention to fairness, equitable participation, and access. Validating on-track/readiness and test security were also added as essential elements. Although the 1999 *Standards* mention “reliability,” the more general term of “precision” is used instead to underscore the need to conceptualize measurement error with other frameworks such as item response theory and generalizability theories. Table 1 presents a brief description of this essential validity evidence. Many of these essential validity elements fall under the validity evidence based on test content (e.g., careful test construction) and internal structure (adequate score reliability, scaling, equating). The types of evidence listed in Table 1 will reemerge when considering the five specific validity sources, which represent the full validity framework. This overlap underscores the fundamental nature of these elements for supporting the use of Smarter Balanced assessments for their intended purposes. Table 1 is followed by a brief description of the potential types of evidence associated with each essential element.

Table 1. Synopsis of Essential Validity Evidence Derived from *Standards* (AERA et al., 1999, p. 17).

Essential Element	Type of Associated Validation Evidence
Careful Test Construction	Examination of test development steps, including construct definition (test specifications and blueprints), item writing, content review, item analysis, alignment studies, and other content validity studies; review of technical documentation such as IRT scaling.
Adequate Measurement Precision (Reliability)	Analysis of test information, conditional standard errors of measurement, generalizability studies, decision accuracy and consistency, and reliability estimates.
Appropriate Test Administration	Review of test administration procedures, including protocols for test irregularities; use of and appropriate assignment of test accommodations.
Appropriate Scoring	Review of scoring procedures (hand-scored, automated), rater agreement analyses, machine/human comparisons (if relevant), generalizability studies, and fairness for subgroups. <i>Test Scoring Specifications</i> (AIR, 2014b).
Accurate Scaling and Equating	Documentation of test design, IRT model choice, scaling and equating procedures, IRT residuals, validating vertical scaling assumptions, third-party verification of horizontal and vertical equating.
Appropriate Standard Setting	Comprehensive standard-setting documentation provided, in Chapter 10, including procedural, internal, and external validity evidence for all achievement-level standards.
Attention to Fairness, Equitable Participation and Access	Review of accommodation policies, implementation of accommodations, sensitivity review, DIF analyses, differential predictive validity analyses, qualitative and statistical analyses of accommodated tests; analysis of participation rates, translations, and other policies.
Validating “On-track/Readiness”	Examining relationships with external variables as well as evidence from internal structure.
Adequate Test Security	Analysis of data integrity policies, test security procedures, monitoring of test administrations, analysis of suspected cheating behavior, item exposure, review of anomalous CAT results.

Careful Test Construction. Validity evidence of careful test construction can derive from a comprehensive inspection of the test development process that reviews all test development

activities. The audit encompasses descriptions of testing purposes, operational definitions of the constructs measured, item development procedures, content reviews, alignment studies, sensitivity and bias reviews, pilot testing, item and DIF analyses, item calibration, item selection, scoring rubrics for constructed-response items, and assembly of tests and clarity of test instructions. For adaptive assessments, the adequacy of the item selection algorithm, particularly in delivering tests that conform to the blueprint, should also be reviewed.

The degree to which the test specifications for the assessment sufficiently reflect the Common Core State Standards and the degree to which the relative weights of the cells in the test specifications reflect the corresponding emphases in the Common Core State Standards should be evaluated (Mislevy & Riconscente, 2006). This entails the use of traditional content validity studies (e.g., Crocker, Miller, & Franks, 1989) and alignment studies (Bhola, Impara, & Buckendahl, 2003; Martone & Sireci, 2009; Porter & Smithson, 2002; Rothman, 2003; Webb, 2007). To evaluate the appropriateness of the test specifications, the process by which the specifications were developed need to be reviewed to ensure that all member states had input and that there was consensus regarding the degree to which the test specifications targeted for the assessment represented the Common Core State Standards. To evaluate the degree to which the summative assessments adequately represent the test specifications requires recruitment and training of qualified and independent subject matter experts in ELA/literacy and mathematics to review the Common Core State Standards in conjunction with the test specifications and Smarter Balanced test items. At least two hypothesized aspects of the assessments can be validated using the content experts. First, the items can be evaluated to ensure that they were appropriately assessing the Common Core of State Standards as intended. Second, the items are measuring the breadth of higher- and lower-order cognitive skills (i.e., Depth of Knowledge) that they are intended to measure. Popham (1992) suggested a criterion of 7 out of 10 subject matter experts (SMEs) rating an item congruent with its standard to confirm that the item fits the standard. Several statistics have been proposed for evaluating item-standard congruence such as Hambleton's (1980) item-objective congruence index and Aiken's (1980) content validity index. In addition, Penfield and Miller (2004) established confidence intervals for subject matter experts mean ratings for content congruence.

Adequate Measurement Precision (Reliability). The notion of measurement precision extends the notion of reliability beyond a descriptive statistic for a test. It refers to the amount of expected variation in a test score or a classification based on a test score. Examples of this type of information include estimates of score reliability, standard errors of measurement, item and test information functions, conditional standard error functions, and estimates of decision accuracy and consistency. Estimates of score reliability typically include internal consistency estimates based on a single test administration (coefficient alpha, stratified alpha, IRT marginal reliability). Generalizability studies that focus on specific facets of measurement are important for identifying the sources of measurement error.

For expected operational adaptive test delivery with multiple content and psychometric constraints, simulations play an important role in evaluating the operational adaptive algorithm and delivery system and the evaluation of measurement error. Test information functions, recovery of simulated examinee ability, and analysis of bias and error are all highly interrelated and can be addressed collectively. All test scores include an error component, the size of which generally varies across test takers. Differences in precision across score ranges are ignored by overall measures of precision that, like test reliability, are aggregated across score levels. However, IRT provides a related pair of test precision measures that are specific to, or conditional on, score level. Both the test information function and the inversely related conditional standard error measure test-precision level across the score scale. (The conditional standard error function is the inverse of the square root of the test information function.) In a simulation environment, the score bias function measures the extent to which score estimates converge to their true values. The smaller the bias and error, the better the

test administration and scoring procedures recover simulated examinee ability. Even if the goal is to measure each student to some fixed criteria for test information/conditional standard error, test precision can vary not just across proficiency levels but also across test takers at the same level of proficiency. Certain students are more easily assessed compared with other ones. Students who respond predictably (as the underlying item-response model expects them to) will be more easily measured by an adaptive test than those who respond in unanticipated ways. Predictable students are well targeted early in a test and are typically presented a series of highly discriminating items. Those students that respond in more unexpected ways will be more difficult to target and less likely to receive informative tests. Much of this inconsistency is unavoidable. However, test administration procedures may differ in the extent to which each test taker is measured on the targeted precision. It should be noted that exceeding the precision target is almost as undesirable as falling short. Measuring some test takers more precisely than necessary wastes resources (in the form of item exposures) that could be used more productively with other test takers.

Appropriate Test Administration. Evidence in this category involves review of test administration manuals and other aspects of the test administration processes. This review can include a review of the materials and processes associated with both standard and accommodated test administrations. Observations of test administrations and a review of proctor and test irregularity reports can be inspected. The policies and procedures for granting and providing accommodations to students with disabilities and English language learners can be reviewed, and case studies of accommodated test administrations should be selected and reviewed to evaluate the degree to which the policies and procedures were followed.

Evidence in this category should also confirm that the routing of test content to students during the linear-on-the-fly and adaptive administration is performing according to expectations and that all computerized scoring programs are accurate. Monitoring of item exposure rates is important as well. The *Standards* (2014 p. 43) point out that one way to evaluate administrations using computer adaptive techniques is to use simulations with known parameters to estimate reliability/precision.

Appropriate Scoring. Validity evidence to confirm that the scoring of Smarter Balanced assessments is appropriate should include a review of scoring documentation. The *Standards* (p. 92) state that such documentation should be presented in sufficient detail and clarity to maximize the accuracy of scoring and the processes for selecting, training, and qualifying scorers. The scoring processes should also include monitoring of all aspects of rater agreement. If any assessments are scored locally, the degree to which the scorers are trained, and the accuracy of their scores, should also be documented. Generalizability studies that quantify various sources of measurement error also provide important evidence, such as characterizing the student by task interactions on performance tasks. For automated scoring, descriptions of the methods used for scoring should be described as well as the development methods that were utilized, such as natural language processing and training and validations studies.

Accurate Scaling and Linking. Scaling and linking are essential activities for producing valid scores and score interpretations for the Smarter Balanced assessments. Scaling activities include item calibration and creation of a standardized scale on which scores are reported. A sound scaling and linking design and representative student samples are critical precursors to conducting the scaling analysis. A sound linking design includes criteria such as content-representative and blueprint-conforming test forms being administered, particularly with respect to common/anchor linking items. Evaluating the adequacy of these scaling and linking activities includes steps that confirm the hypothesized dimensionality of the assessments and the viability of a single construct (dimension) across grades, the performance of different IRT scaling models, scrutiny of the linking results, and potentially examining the invariance of the equating across subgroups of students. A major

assumption for the use of more traditional scaling methods is that a given test is essentially unidimensional, consisting of a major dimension along with some minor ones. The nature of the change over grade levels is characterized in the common items given across grade levels that are used to construct the vertical scale. The viability of a vertical scale depends on this major dimension being consistent across levels of the test. The influence of the minor dimensions will determine how the construct shifts across grade levels. Since the ELA/literacy and mathematics assessments were vertically linked across grades, evidence concerning the nature of change in the construct over levels of the test and its plausibility are necessary. A “cross validation study,” where an independent third party replicates the scaling and linking, provides an important validity check on the accuracy of the equating. Once the calibrated item pool is available, a choice of IRT scoring methods is necessary, such as maximum likelihood estimation (Thissen & Wainer, 2001), that forms the basis for achievement level reporting.

Appropriate Standard Setting. When achievement-level (i.e., proficiency) standards are set on tests, scale scores often become less important than the proficiency classifications students receive which are the central focus of many accountability systems. There are many different methods for setting standards, but regardless of the method used, there must be sufficient validity evidence to support the classification of students into achievement levels. The Smarter Balanced summative assessments used achievement levels, some of which signified “on track” to “college readiness” (grades 3-8) or “college ready” (grade 11). An additional element was the articulation of cut points across grade levels in the context of a vertical scale. The primary assumption here is that the cut points increase across grade levels in a logical progression that reflects increased levels of achievement that ultimately culminate in “readiness” in grade 11. Articulated achievement levels means the proficiency cut scores maintain some consistent level of stringency or pattern across grades.

Gathering and documenting validity evidence for standards set on educational tests can be categorized into three categories—procedural, internal, and external (Kane, 1994; 2001). Procedural evidence for standard setting “focuses on the appropriateness of the procedures used and the quality of the implementation of these procedures” (Kane, 1994, p. 437). The selection of qualified standard-setting panelists, appropriate training of panelists, clarity in defining the tasks and goals of the study, appropriate data collection procedures, and proper implementation of the method are all examples of procedural evidence. Internal evidence for evaluating standard-setting studies focuses on the expected consistency of results if the study was replicated. A primary criterion is the standard error of the cut score. However, calculation of this standard error is difficult due to dependence among panelists’ ratings and practical factors (e.g., time and expense in conducting independent replications). Oftentimes, evaluations of the variability across panelists within a single study and the degree to which this variability decreases across subsequent rounds of the study are presented as internal validity evidence. However, as Kane (2001) pointed out:

A high level of consistency across participants is not to be expected and is not necessarily desirable; participants may have different opinions about performance standards. However, large discrepancies can undermine the process by generating unacceptably large standard errors in the cut scores and may indicate problems in the training of participants (p. 73).

In addition to simply reporting the standard error of the cut score, Kane (2001) suggested that consistency can be evaluated across independent panels, subgroups of panelists, or assessment tasks (e.g., item formats), or by using generalizability theory to gauge the amount of variability in panelists’ ratings attributed to these different factors. Another source of internal validity evidence proposed by Kane was to evaluate the performance of students near the cut score on specific items to see if their performance was consistent with the panelists’ predictions. External validity evidence

for standard setting involves studying the degree to which the classifications of students based on test scores are consistent with other measures of their achievement in the same subject area. External validity evidence includes classification consistency across different standard-setting methods applied to the same test, to tests of mean differences across examinees classified in different achievement levels on other measures of achievement, and the degree to which external ratings of student performance are congruent with their test-based achievement-level classifications. External validity evidence is particularly important for validating the “college and career readiness” standards set on the summative assessments. A number of measures for determining college readiness already exists. The degree to which the constructs measured by these external assessments overlap with the Smarter Balanced summative assessments and the degree to which their definitions of readiness are similar (or different) should be addressed by Smarter Balanced.

Attention to Fairness, Equitable Participation, and Access. Chapter 3 of the *Standards* (p. 49-70) addresses fairness in testing. The intent of the Smarter Balanced system is to provide additional flexibility and remove construct-irrelevant barriers that prevent students from taking the test or demonstrating their best performance. Construct irrelevant barriers can be minimized through test design and testing adaptations. Evidence-centered design, item specifications, usability, accessibility, and accommodations guidelines, bias and sensitivity guidelines, and reviews by content developers are all used to develop items and tasks that ensure the targeted constructs are measured accurately. A critical aspect of access is the ability to deliver items, tasks, and the collection of student responses in a way that maximizes validity for each student. Equitable participation and access ensures that all students can take the test in a way that allows them to comprehend and respond appropriately. This includes, but is not limited to, English Language Learners (ELLs), students with disabilities, and ELLs with disabilities. The *Standards* also specify an aspect of fairness as a lack of measurement bias. Characteristics of items that are construct irrelevant can affect the performance by members of some identifiable subgroups, which is called differential item functioning. Many methods exist for investigating differential item functioning statistically. For an item exhibiting DIF, additional investigation is required in order to conclude it is biased.

Validating “On-Track/Readiness” “On-track” denotes notions concerning expectations and adequate levels of growth being demonstrated. Studies related to expected growth will be conducted after two iterations of operational testing have been conducted. States use a variety of growth models, and the Consortium is not recommending or discouraging any specific model. Growth is defined as improvement in performance for a given group of students over time—such as improvement from grade 6 to grade 7. Vertical scales can facilitate the measurement of student growth and permit direct comparisons of change using scale scores (or status) across different grades or change within a year.

College and career readiness may have substantial overlap since employers and colleges have similar perspectives on the level of knowledge and skills required for entry (Achieve, 2004; ACT, 2006). However, others have argued the benchmarks for college and career readiness will be very different (Camara, 2013; Loomis, 2011). On-track for college readiness implies the acquisition of knowledge and the mastery of specific skills deemed important as students progress through elementary, middle, and high school that are stipulated in the Common Core State Standards. Validity studies such as content alignment can be used to confirm that the Smarter Balanced assessments are targeting the correct Common Core State Standards and adequately represent these standards. However, studies of this type do not confirm that the Common Core State Standards actually contain the appropriate knowledge and skills to support college and career readiness (Sireci, 2012). At the higher education level, Conley, Drummond, de Gonzalez, Rooseboom, and Stout (2011) conducted a national survey of postsecondary institutions to evaluate the degree to which the grade 11 Common Core State Standards contain the knowledge and skills

associated with college readiness. They found that most of approximately 2,000 college professors rated the Common Core State Standards as highly important for readiness in their courses. Similarly, Vasavada, Carman, Hart, and Luisser (2010) found strong alignment between College Board assessments of college readiness and the Common Core State Standards. The additional evidence required for readiness is evidence that these standards reflect the appropriate prerequisite skills in mathematics and ELA/literacy that are needed to bypass remedial college courses and to successfully begin postsecondary education or a career. Other validity evidence based on test content is the content overlap (alignment) studies that will be undertaken to gauge the similarity of knowledge and skills measured across the Summative assessments and external assessments that are used to evaluate the readiness standards (Sireci, 2012). Postsecondary admissions tests (e.g., ACT, SAT) and college placement tests (e.g., Accuplacer, Advanced Placement, & Compass) can be used in concurrent and predictive validity studies. This requires the overlap in the skills measured to be identified to derive the proper inferences.

The degree to which other measures of college readiness benchmarks are consistent with the Smarter Balanced readiness standards can be examined. Camara (2013) listed seven criteria that have been or could be used for setting or evaluating college readiness benchmarks on the Smarter Balanced assessments. These criteria are:

- persistence to second year;
- graduation or completion of a degree or certification program;
- time to degree completion (e.g., six years to earn a bachelor's degree);
- placement into college credit courses;
- exemption from remediation courses;
- college grades in specific courses; and
- college grade-point average.

Validity evidence based on relations to other external variables for the purpose of classifying students as college ready can involve both correlation type studies and classification consistency analyses.

The college and career readiness standard is intentionally integrated with the “on-track” standards set at the lower grade levels with the intended consequence that the system better prepares students for college or careers by the time they graduate high school. These college and career readiness outcomes can be appraised using trends in college completion and remedial course enrollments over time, and by surveying secondary and postsecondary educators about students’ proficiencies. The recommended studies based on test consequences for college and career readiness purposes should include teacher surveys regarding changes in student achievement and preparedness over time and changes in their instruction over time. Students can be surveyed regarding college and career aspirations. Student and teacher samples that are representative at the state level would suffice for these studies. Validity evidence based on the consequences of the college and career readiness standard should involve analysis of secondary and postsecondary enrollment and persistence, changes in course-taking patterns over time, and teacher retention for teachers in mathematics and ELA/literacy.

Studies of the relationship of Smarter scores to college course enrollment, grades and course completion will be conducted as students using the Consortium tests enter college. The Consortium has created career cluster readiness frameworks Smarter Balanced, (2013) to inform alignment of test content to career-related skills and to aid in score interpretation.

Adequate Test Security. Test security is a prerequisite to validity. As described by NCME (2012), “When cheating occurs, the public loses confidence in the testing program and in the educational system which may have serious educational, fiscal, and political consequences.” Threats to test security include cheating behaviors by students, teachers, or others who have unwarranted access to testing materials. A lack of test security may result in the exposure of items before tests are administered, students copying or sharing their answers, or changing students’ answers to test questions in fixed-form tests. Many proactive steps can be taken to reduce, eliminate, and evaluate cheating. The first step is to keep confidential test material secure and have solid procedures in place for maintaining the security of paper and electronic materials. The NCME (2012) document on data integrity outlined several important areas of test security. These areas include procedures that should be in place before, during, and after testing. The activities prior to testing include securing the development and delivery of test materials. During testing, activities include adequate proctoring to prevent cheating, imposters, and other threats. After testing, checking social media for item content and the forensic analysis of students’ responses and answer changes and aberrant score changes over time are also beneficial. The goal of these security activities is to ensure that test data are “free from the effects of cheating and security breaches and represent the true achievement measures of students who are sufficiently and appropriately engaged in the test administration” (NCME, 2012, p. 3).

The evaluation of the test security procedures of the assessments involved a review of the test security procedures and data forensics by Smarter Balanced. The NCME (2012) document on test data integrity suggests that security policies should address the following:

... staff training and professional development, maintaining security of materials and other prevention activities, appropriate and inappropriate test preparation and test administration activities, data collection and forensic analyses, incident reporting, investigation, enforcement, and consequences. Further, the policy should document the staff authorized to respond to questions about the policy and outline the roles and responsibilities of individuals if a test security breach arises. The policy should also have a communication and remediation response plan in place (if, when, how, who) for contacting impacted parties, correcting the problem and communicating with media in a transparent manner (p. 4).

With adaptive test administration, the probability of students receiving the same items at similar times is low, and the probability of answer copying is very low. However, consistent with other CAT programs, item exposure rates should be carefully monitored on an ongoing basis. Rules for rotating items out of the summative assessment with comparatively high exposure rates are needed. Due to the nature of performance tasks that are more memorable and subject to practice effects, they will need to be replaced or transitioned frequently.

Summary of Essential Validity Evidence based on the Smarter Pilot- and Field Tests

Other chapters of the Smarter Balanced technical report describe the evidence and studies performed to date for the Smarter Balanced Assessment Field Test. Tables 2 to 10 list the essential validity elements and associated evidence types for each one. For example, Table 2 presents the essential validity element for “Careful Test Construction. It lists the types of validation evidence associated with that element in terms of a short label, and provides the associated evidence source. For the evidence source, the chapter and section in parenthesis refers to this Technical Report. When appropriate, other Smarter Balanced documentation or reports are listed in italics. The reader will need to make a judgment as to the importance of the essential validity element presented and the number, quality, and types of supporting evidence.

Table 2. Essential Validity Evidence for the Summative and Interim Assessments for Careful Test Construction.

Evidence Type	Evidence Source
Theory of Action/testing purposes clearly stated	<i>Smarter Balanced Theory of Action</i> , Introduction (Theory of Action)
Evidence centered design implemented	Test Design, (Evidence Centered Design), <i>Smarter Balanced Bibliography</i> , <i>General Item Specifications</i>
Test specifications sufficiently documented	Test Design, (Operational Summative Assessment Blueprints and Specifications), <i>Performance Task Specification</i> , <i>Mathematics Performance Task Specifications</i>
Construct definition	Test Design, (Operational Summative Assessment Blueprints and Specifications), <i>ELA/literacy Content Specifications</i> , <i>Mathematics Content Specifications</i>
Item writers appropriately recruited and trained	<i>Mathematics Performance Task Specifications</i>
Items adhere to item writing style guidelines	<i>General Item Specifications</i>
Items reviewed for content quality and technical adequacy	Field Test Data Step and Classical Test Analysis (Item Flagging Criteria for Content Data Review)
Content validity/alignment studies	<i>General Item Specifications</i> ; <i>Smarter Balanced Assessment Consortium Alignment Study</i> , <i>HumRRO</i>
Sensitivity reviews	Test Fairness, (Definitions for Validity, Bias, Sensitivity, and Fairness), <i>Standards for Educational and Psychological Testing</i> , <i>Smarter Balanced Bias and Sensitivity Guidelines</i> , <i>ETS Guidelines for Fairness Review of Assessments</i>
Test booklets conform to test blueprints	Field Test Design, Sampling, and Administration (Numbers and Characteristics of Items and Students Obtained in the Field Test), Field Test Design, Sampling, and Administration (Field Test Delivery Modes), <i>Smarter Balanced Adaptive Item Selection Algorithm Design Report</i> , <i>General Item Specifications</i> , <i>Simulations studies from AIR and CRESST</i> .
Data review (Classical)	Pilot Study, (Pilot Classical Test Results); Field technical review.
Item selection/delivery based on content criteria	Test Design, (Operational Summative Assessment Blueprints and Specifications), Field Test Design,

Evidence Type	Evidence Source
	Sampling, and Administration, (Field Test Delivery Modes), <i>Adaptive Selection Algorithm</i> (Cohen & Albright, 2014)
IRT Item calibration	Pilot Test, (Dimensionality Study), Pilot Test, (IRT Model Comparison), Field Test IRT Scaling and Linking Analyses, (All sections)

Table 3. Essential Validity Evidence for the Summative and Interim Assessments for Adequate Measurement Precision (Reliability).

Evidence Type	Evidence Source
Test reliability (Internal Consistency)	Data Step and Classical Test Analysis (Field Test Results); Simulation studies by AIR and CRESST.
IRT item fit	Field Test IRT Scaling and Linking Analyses (Horizontal and Vertical Scaling Results), Pilot Test, (Item Response Theory [IRT] Model Comparison)
Conditional standard error of measurement (CSEM) for ability	Field Test IRT Scaling and Linking Analyses (Horizontal and Vertical Scaling Results) Simulation studies by AIR and CRESST.
Standard error of measurement (Classical)	Data Step and Classical Test Analysis (Field Test Results)
IRT test information	Field Test IRT Scaling and Linking Analyses (Horizontal and Vertical Scaling Results)
Generalizability studies	
Cut-score decision consistency and accuracy	Simulation studies by AIR and CRESST.

Table 4. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Test Administration.

Evidence Type	Evidence Source
Availability of training and practice modules	<i>Test Administration Manual</i>
Clearly defined instructions	<i>Test Administration Manual, Field Test Design, Sampling and Administration, (Field Test Administration and Security), Smarter Balanced Technical Specifications Manual, Calculator Availability Information for 2014 Field Test</i>
Test delivery system functioned as expected	<i>Smarter Balanced “Tests of the Test” Successful: Field Test Provides Clear Path Forward; Simulation studies by AIR and CRESST.</i>

Table 5. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Scoring

Evidence Type	Evidence Source
Sufficient levels of rater agreement	<i>Pilot Test: (Evaluation of Reliability and Validity for Automated Scoring Models in the Pilot), Field Test: Automated Scoring Research Studies</i>
Scoring rubrics for constructed-response items are reviewed	<i>Smarter Balanced Scoring Guide for Selected Short-Text Mathematics Items (Field Test 2014), Performance Tasks Specifications</i>
Adaptive item selection algorithm documented	<i>Smarter Balanced Adaptive Item Selection Algorithm Design Report (Cohen & Albright, 2014)</i>
Content conforming tests are delivered	Field Test Design, Sampling and Administration (Numbers and Characteristics of Items and Students Obtained in the Field Test)
Rationale, development, and rater agreement for automated scoring	<i>Smarter Balanced Pilot Automated Scoring Research Studies, Field Test: Automated Scoring Research Studies</i>

Table 6. Essential Validity Evidence for the Summative and Interim Assessments for Accurate Scaling and Linking.

Evidence Type	Evidence Source
Sample is representative	Field Test Design, Sampling and Administration, (Sampling Results)
Rationale for IRT model choice is provided	Pilot Test (Dimensionality Study), Pilot Test (Item Response Theory Model Comparison)
Calibration and linking design is appropriate	Field Test IRT Scaling and Linking Analyses, (Vertical Scaling: Linking Across Multiple Grades), Field Test Design, Sampling and Administration, (Field Test), Field Test Design, Sampling and Administration (Field Test Student Sampling Design)
Accurate IRT horizontal and vertical scaling met	Field Test IRT Scaling and Linking Analyses, (All Sections)
Accurate equating methods applied	Field Test Design, Sampling and Administration, (Field Test), Field Test Design, Sampling, and Administration (Field Test Student Sampling Design), Field Test IRT Scaling and Linking Analyses, (Assumptions and Interpretive Cautions Concerning Vertical Scales) Field Test IRT Scaling and Linking Analyses, (Horizontal and Vertical Scaling Results)
Assumptions for establishing vertical scale are met	Field Test IRT Scaling and Linking Analyses, (Assumptions and Interpretive Cautions Concerning Vertical Scales), Field Test IRT Scaling and Linking Analyses, (Vertical Linking Procedures)

Table 7. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Standard Setting.

Evidence Type	Evidence Source
Justification of standard setting method(s)	Achievement Level Setting (The Bookmark Procedure), <i>Achievement Level Setting Plan</i>
Panelist recruitment and training	Achievement Level Setting (Recruitment and selection of panelists)
Clarity of goals/tasks	<i>Achievement Level Setting Plan</i>
Clear achievement level descriptors	<i>Initial Achievement Level Descriptors and College Content-Readiness Policy, Achievement Level Setting (Achievement Level Descriptors), Interpretation and Use of Scores and Achievement Levels</i>
Appropriate data collection	Achievement Level Setting (The Bookmark Procedure), <i>Achievement Level Setting Plan</i>
Implementation	Achievement Level Setting (All Sections), <i>Achievement Level Setting Plan</i>
Panelist confidence	Achievement Level Setting (Round-by-round item review and discussion)
Sufficient documentation	Achievement Level Setting (All Sections), <i>Achievement Level Setting Plan</i>
Sufficient inter-panelist consistency	Achievement Level Setting (Round-by-round item review and discussion)
Across grade level articulation	Achievement Level Setting (Design and Implementation of the Cross-Grade Review Committee)
Reasonableness of achievement standards	Achievement Level Setting (All Sections), <i>Statements of Support: Achievement Level Setting, Achievement Level Descriptors and College Content-Readiness, Achievement Level Setting Statements of Support</i>

Table 8. Essential Validity Evidence for the Summative and Interim Assessments for Attention to Fairness, Equitable Participation and Access.

Evidence Type	Evidence Source
DIF Analysis	Field Test Datastep and Classical Test Analysis (Differential Item Functioning (DIF) Analyses for the Calibration Item Pool)
Equitable Participation	<i>Usability, Accessibility, and Accommodations Guidelines</i>
Universal Design, Assessment Supports, and Accommodations	<i>Usability, Accessibility, and Accommodations Guidelines; Test Fairness (Usability, Accessibility, and Accommodations Guidelines: Intended Audience and Recommended Applications), General Item Specifications, Signing Guidelines, Tactile Accessibility Guidelines</i>
Support for English Language Learners	<i>Guidelines for Accessibility for English Language Learners</i>
Bias and Sensitivity	<i>Smarter Balanced Assessment Consortium: Bias and Sensitivity Guidelines, Test Fairness (Definitions for Validity, Bias, Sensitivity, and Fairness)</i>

Table 9. Essential Validity Evidence for the Summative and Interim Assessments for Validating “On-Track/Readiness”.

Evidence Type	Evidence Source
Relationships with External Variables/Tests	Field Test Design, Sampling, and Administration (Linking PISA and NAEP to Smarter Balanced Assessments), Data Step and Classical Test Analysis (Field Test Results)
Operational definition of college content-readiness	<i>Study of the Relationship Between the Early Assessment Program and the Smarter Balanced Field Tests, ELA/literacy Achievement Level Descriptors and College Content-Readiness Policy, Mathematics Achievement Level Descriptors and College Content-Readiness Policy, Reaching the Goal: The Applicability and Importance of the Common Core State Standards to College and Career Readiness</i>

Table 10. Essential Validity Evidence for the Summative and Interim Assessments for Adequate Test Security.

Evidence Type	Evidence Source
Test security procedures and exceptions escalation documented	<i>Online Field Test Administration Manual for Spring 2014 Field Tests of English Language Arts/Literacy and Mathematics</i> , Field Test Design, Sampling, and Administration, (Field Test Administration and Security), <i>The Smarter Balanced Technology Strategy Framework and Testing Device Requirements</i>

The *Standards'* Five Primary Sources of Validity Evidence

The five sources of validity evidence serve as organizing principles and represent a comprehensive framework for evaluating validity for Smarter Balanced. These sources of validity evidence are intended to emphasize different aspects of validity. However, since validity is a unitary concept, they do not constitute distinct types of validity. These five sources of validity evidence consist of (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing. They are briefly described below:

1. Validity evidence based on *test content* refers to traditional forms of content validity evidence, such as the rating of test specifications and test items (Crocker, Miller, & Franks, 1989; Sireci, 1998), as well as “alignment” methods for educational tests that evaluate the interactions between curriculum frameworks, testing, and instruction (Rothman, Slattery, Vranek, & Resnick, 2002; Bholá, Impara & Buckendahl, 2003; Martone & Sireci, 2009). The degree to which (a) the Smarter Balanced test specifications captured the Common Core State Standards and (b) the items adequately represent the domains delineated in the test specifications, were demonstrated in the alignment studies. The major assumption here is that the knowledge, skills, and abilities measured by the Smarter Balanced assessments are consistent with the ones specified in the Common Core State Standards. Administration and scoring can be considered as aspects of content-based evidence. With computer adaptive testing, an extra dimension of test content is to ensure that the tests administered to students conform to the test blueprint.
2. Validity evidence based on *response processes* refers to “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA et al., 1999 p. 12). This evidence might include documentation of such activities as
 - interviewing students concerning their responses to test items (i.e., speak alouds);
 - systematic observations of test response behavior;
 - evaluation of the criteria used by judges when scoring performance tasks, analysis of student item-response-time data, features scored by automated algorithms; and
 - evaluation of the reasoning processes students employ when solving test items (Emberetson, 1983; Messick, 1989; Mislevy, 2009).

This type of evidence was used to confirm that the Smarter Balanced assessments are measuring the cognitive skills that are intended to be the objects of measurement and that students are using these targeted skills to respond to the items.

3. Validity evidence based on *internal structure* refers to statistical analyses of item and score subdomains to investigate the primary and secondary (if any) dimensions measured by an assessment. Procedures for gathering such evidence include factor analysis or multidimensional IRT scaling (both exploratory and confirmatory). With a vertical scale, a consistent primary dimension or construct shift across the levels of the test should be maintained. Internal structure evidence also evaluates the “strength” or “salience” of the major dimensions underlying an assessment using indices of measurement precision such as test reliability, decision accuracy and consistency, generalizability coefficients, conditional and unconditional standard errors of measurement, and test information functions. In addition, analysis of item functioning using Item Response Theory (IRT) and differential item functioning (DIF) fall under the internal structure category. For Smarter Balanced, a dimensionality study was conducted in the Pilot Test to determine the factor structure of the assessments and the types of scales developed as well as the associated IRT models used to calibrate them.
4. Evidence based on *relations to other variables* refers to traditional forms of criterion-related validity evidence such as concurrent and predictive validity, as well as more comprehensive investigations of the relationships among test scores and other variables such as multitrait-multimethod studies (Campbell & Fiske, 1959). These external variables can be used to evaluate hypothesized relationships between test scores and other measures of student achievement (e.g., test scores and teacher grades), the degree to which different tests actually measure different skills and the utility of test scores for predicting specific criteria (e.g., college grades). This type of evidence is essential for supporting the validity of certain inferences based on scores from the Smarter Balanced assessments for certifying college and career readiness, which is one of the primary test purposes. A subset of students who took NAEP and PISA items also took Smarter Balanced items and performance tasks. A summary of the resulting item performance for NAEP, PISA, and all Smarter Balanced items was conducted.
5. Finally, evidence based on *consequences of testing* refers to the evaluation of the intended and unintended consequences associated with a testing program. Examples of evidence based on testing consequences include investigations of adverse impact, evaluation of the effects of testing on instruction, and evaluation of the effects of testing on issues such as high school dropout rates. With respect to educational tests, the *Standards* stress the importance of evaluating test consequences. For example, they state,

When educational testing programs are mandated . . . the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences.

Consequences resulting from the use of the test, both intended and unintended, should also be examined by the test user (AERA et al., 1999, p. 145).

Investigations of testing consequences relevant to the Smarter Balanced goals include analyses of students’ opportunity to learn with regard to the Common Core State Standards, and analyses of changes in textbooks and instructional approaches. Unintended consequences, such as changes in instruction, diminished morale among teachers and students, increased pressure on students leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging, can be evaluated.

Purposes of the Smarter Balanced System for Summative, Interim, and Formative Assessments

The Smarter Balanced purpose statement refers to three categories consisting of summative, interim, and formative assessment resources. To derive the statements of purpose listed below, panels consisting of Smarter Balanced leadership, including the Executive Director, Smarter Balanced staff, Dr. Stephen Sireci and key personnel from Consortium states were convened.

The purposes of the Smarter Balanced summative assessments are to provide valid, reliable, and fair information concerning:

- 1) Students' ELA/literacy and mathematics achievement with respect to those CCSS measured by the ELA/literacy and mathematics summative assessments.
- 2) Whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on track for achieving college readiness.
- 3) Whether grade 11 students have sufficient academic proficiency in ELA/literacy and mathematics to be ready to take credit-bearing college courses.
- 4) Students' annual progress toward college and career readiness in ELA/literacy and mathematics.
- 5) How instruction can be improved at the classroom, school, district, and state levels.
- 6) Students' ELA/literacy and mathematics proficiencies for federal accountability purposes and potentially for state and local accountability systems.
- 7) Students' achievement in ELA/literacy and mathematics that is equitable for all students and subgroups of students.

Providing valid, reliable, and fair information about students' ELA/literacy and mathematics achievement with respect to the Common Core State Standards as measured by the summative assessments is central. Validity evidence to support this purpose derives from at least three sources—test content, internal structure, and response processes. With respect to test content, evidence confirming that the content of the assessments adequately represents the Common Core State Standards to be measured in each grade and subject area is essential. Content domain representation and congruence to the Common Core State Standards must be substantiated. Validity evidence based on internal structure involves analysis of item response data to confirm that the dimensionality of the data matches the intended structure and supports the scores that are reported. Measures of reliability, test information, and other aspects of measurement precision are also relevant. Validity evidence based on response processes should confirm that the items designed to measure higher-order cognitive skills are tapping into those targeted skills.

The Smarter Balanced assessments focus on the provision of valid, reliable, and fair information concerning whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on-track for achieving college readiness. Secondly, the intent is to provide valid, reliable, and fair information about whether grade 11 students have sufficient academic proficiency in ELA and mathematics to be ready to take credit-bearing college courses or are career-ready. These two purpose statements reflect that the Smarter Balanced summative assessments will be used to classify students into achievement levels. Before grade 11, one achievement level will be used at each grade to signal whether students are “on-track” to college or career readiness. At grade 11, the achievement levels will include a “college and career readiness” category. These classification decisions require validation that can be derived from four sources—test content, internal structure, relations with external variables, and testing consequences.

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

- 1) Student progress toward mastery of the skills in ELA/literacy and mathematics measured by the summative assessment.
- 2) Student performance at the claim or cluster of Assessment Targets so teachers and administrators can track student progress throughout the year and adjust instruction accordingly.
- 3) Individual and group (e.g., school, district) performance at the claim level in ELA/literacy and mathematics to determine how teaching and learning can best be targeted.
- 4) Student progress toward the mastery of skills measured in ELA/literacy and mathematics across all students and subgroups of students.

The Smarter Balanced interim assessments differ from the summative assessments in that they are optional, non-secure components that can be administered multiple times within a school year and are designed to provide information at a finer level of detail with respect to students' strengths and weaknesses in relation to the Common Core State Standards. The interim assessments are intended to help teachers focus assessment on the most relevant aspects of classroom instruction at a particular point in time. They are also intended to play a role in professional development, particularly in cases in which teachers can determine how scoring rubrics align with the content standards and have the opportunity to score student responses to items.

The purposes of the Smarter Balanced formative assessment resources are to provide measurement tools and resources to:

- 1) Improve teaching and learning.
- 2) Monitor student progress throughout the school year.
- 3) Help teachers and other educators align instruction, curricula, and assessment.
- 4) Assist teachers and other educators in using the summative and interim assessments to improve instruction at the individual student and classroom levels.
- 5) Illustrate how teachers and other educators can use assessment data to engage students in monitoring their own learning.

The Formative Assessment Resources are not assessments per se, and so the evidence in support of their intended purposes extends beyond the five sources of validity evidence and requires a program evaluation approach. Tables 11 and 12 illustrate the validation framework for the summative and interim assessments by cross-referencing the purpose statements for each component with the five sources of validity evidence. The check marks in the cells indicate the type of evidence that could be used for validating a specific purpose. While this presentation is general, it is useful for understanding which sources of validity evidence are most important for specific test purposes. For example, for purposes related to providing information about students' knowledge and skills, validity evidence based on test content is critical. For purposes related to classifying students into achievement categories such as "on-track" or "college-ready", validity evidence based on internal structure is needed since evidence of this type also relies on sufficient level of decision consistency and accuracy being demonstrated.

Table 11. Validity Framework for Smarter Balanced Summative Assessments.

Purpose	Source of Validity Evidence				
	Content	Internal Structure	Relations with External Variables	Response Processes	Test Consequences
Report achievement with respect to the CCSS* as measured by the ELA/literacy and mathematics summative assessments	✓	✓	✓	✓	
Assess whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on track for college readiness	✓	✓	✓		✓
Assess whether grade 11 students have sufficient academic proficiency in ELA/literacy and mathematics to be ready to take credit-bearing college courses	✓	✓	✓		✓
Measure students' annual progress toward college and career readiness in ELA/literacy and mathematics	✓	✓	✓		✓
Inform how instruction can be improved at the classroom, school, district, and state levels	✓				✓
Report students' ELA/literacy and mathematics proficiency for Federal accountability purposes and potentially for state and local accountability systems	✓	✓	✓		✓
Assess students' achievement in ELA/literacy and mathematics in a manner that is equitable for <i>all</i> students and subgroups of students	✓	✓	✓	✓	✓

Note: *CCSS denotes Common Core State Standards.

Table 12. Validity Framework for Smarter Balanced Interim Assessments.

Purpose	Source of Validity Evidence				
	Content	Internal Structure	Relations with External Variables	Response Processes	Test Consequences
Assess student mastery of the skills and knowledge measured in ELA/literacy and mathematics	✓	✓		✓	
Assess students' performance at the claim level or finer so teachers and administrators can track student progress throughout the year and adjust instruction accordingly	✓	✓			✓
Assess individual and group (e.g., school, district) performance at the claim level in ELA/literacy and mathematics to determine whether teaching and learning are on target	✓	✓	✓		✓
Measure student progress toward the mastery of skills measured in ELA/literacy and mathematics across all subgroups	✓	✓	✓	✓	✓

Evidence Using the Five Primary Sources of Validity Framework

Table 13 lists the evidence type, the associated Smarter Balanced evidence demonstrated to date (or not), and the relevancy of the *Standards* five primary validity evidence sources. It further cross-classifies each piece of evidence with the Smarter Balanced Summative, Interim, and Formative Test Purposes. The evidence demonstrated lists the relevant chapter of the Technical Report or the relevant external documents. Even if several sources of evidence are presented, the reader must still make a judgment whether sufficient supporting evidence has been offered. For instance, the reader may question if, for vertical scaling, there is an increase in the difficulty of the assessments as the grade level increases, with generally greater student proficiency demonstrated in higher grades relative to lower grades. In the case where the evidence source in the table is blank simply means no evidence is available to date. The full complement of validity evidence is ambitious in both its scope and the resources required to fulfill it. The table is also useful in that it serves to identify current gaps in the ongoing validity argument and identify the sorts of evidence needed going forward, as proposed in the Smarter Balanced comprehensive research agenda (Sireci, 2012).

Table 13. Listing of Evidence Type, Evidence Source, and Primary Validity Source for Summative, Interim, and Formative Test Purposes.

Evidence Type	Evidence Source	Primary Validity Source	Summative							Interim				Formative							
			1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5			
Content validity and alignment	Test Design	1	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	
Sensitivity and bias review	Test Design	1	✓					✓	✓					✓							
Evidence Centered Design	Test Design, <i>General Item Specifications</i>	1, 2	✓	✓	✓	✓					✓	✓	✓								
Subdomain scores (e.g., claims)	Test Design	1, 3										✓	✓								
Scoring (raw scores)	Datastep and Classical Test Analysis	1, 3	✓	✓	✓	✓			✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓
Standard setting	Achievement Level Setting	1, 3, 4	✓	✓	✓	✓	✓	✓	✓				✓			✓			✓		
Test construction practices	Test Design	1, 3	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓				✓	✓	✓	
Fairness	Test Fairness	1, 2, 3, 4, 5	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	
Scope and sequence of curriculum		1, 5					✓					✓	✓			✓	✓	✓	✓	✓	
Test administration	Field Test Design, Sampling, and Administration	1, 5	✓						✓					✓			✓				

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative					
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	
Equitable participation & access	Test Fairness, Field Test Design, Sampling, and Administration, and <i>Usability, Accessibility, and Accommodations Guidelines</i>	1, 5	✓						✓				✓					✓	✓
Test accommodations	Test Design, Test Fairness, <i>Usability, Accessibility, and Accommodations Guidelines</i>	1, 5	✓						✓	✓			✓						
Formative resources development and implementation		1, 5													✓	✓	✓	✓	✓
Cognitive skills, think-aloud protocols		2	✓				✓	✓			✓								
Item response time		2	✓				✓	✓			✓								
Horizontal and vertical scales	Field Test IRT Scaling and Linking Analyses	3		✓	✓	✓		✓	✓		✓								
Decision consistency and accuracy	Achievement Level Setting	3		✓	✓	✓		✓	✓		✓		✓						

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative					
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	
IRT fit analysis	Field Test IRT Scaling and Linking Analyses	3		✓		✓		✓	✓		✓								
Reliability and standard error estimation	Field Test IRT Scaling and Linking Analyses	3	✓	✓	✓	✓	✓	✓	✓		✓	✓							
	and Datastep and Classical Test Analysis																		
Reliability of aggregate statistics		3	✓					✓	✓				✓				✓		
Generalizability studies		3		✓	✓	✓		✓											
Item parameter drift		3		✓	✓	✓					✓								
Test dimensionality	Pilot Test	3	✓			✓	✓				✓	✓	✓						
CAT algorithm		2, 3		✓	✓							✓	✓						
Mode comparability		3			✓														
Automated scoring	Pilot Test, <i>Field Test: Automated Scoring Research Studies</i>	3	✓								✓								

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative						
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5		
Invariance of test structure		3	√					√	√					√						
Test security	Field Test Design, Administration, and Sampling	3, 4	√	√	√			√			√									
Convergent/discriminant validity		3, 4	√					√			√	√	√							
Differential item functioning	Test Fairness	3, 5	√					√					√							
Sensitivity to instruction		4	√	√	√	√	√	√	√		√	√	√	√						
Criterion-related validation of on-track		4		√																
Criterion-related studies of change in achievement/growth		4		√		√		√	√		√	√	√							
Criterion-related validation of readiness		4		√	√	√		√			√									
Differential predictive validity		4	√					√	√					√						
Group differences	Test Fairness, Datastep and Classical Test Analysis	4		√	√	√		√	√					√						
Classroom artifacts		4, 5					√					√				√	√	√	√	√

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative					
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	
Perspective of postsecondary educators		5			✓	✓													
College enrollment, dropout, courses taken		5					✓		✓				✓						
Teacher morale/perception of test utility		5					✓	✓			✓	✓	✓	✓		✓		✓	✓
Teacher perception on changes in student learning		5		✓	✓	✓	✓		✓		✓	✓	✓	✓		✓	✓		✓
Student perspective		5			✓			✓											
Educator interviews, and focus groups		5		✓	✓														
Score report utility and clarity		5	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓
Score report usage rates		5					✓								✓	✓	✓	✓	✓
Follow-up on specific student decisions		5		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓		✓
Interim usage statistics		5									✓	✓	✓	✓					
High efficacy users of interim assessments		5									✓	✓	✓	✓					
Formative usage statistics		5														✓	✓	✓	✓
Collaborative leadership		5														✓		✓	✓

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative					
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	
Utility of formative assessment		5													✓	✓	✓	✓	✓
Formative assessment student perception		5													✓	✓	✓	✓	✓
Parent perception of formative assessment		5													✓				✓
Critique of Theory of Action		5	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Comparison with NAEP, PISA, TIMSS		1, 3, 4, 5																	
Summary of validity evidence supporting seven Theory of Action principles		5	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓

Note: Primary Validity Source: 1=Test Content, 2=Response Processes, 3=Internal Structure, 4=Relations to other variables, 5=Testing consequences

Conclusion for Field Test Validity Results

Validation is an ongoing, essentially perpetual endeavor in which additional evidence can be provided but one can never absolutely “assert” an assessment is perfectly valid (Haertel, 1999). This is particularly true for the many purposes typically placed on tests. Program requirements are often subject to change and the populations 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. A review of the purpose statements and the available validity evidence determines the degree to which the principles outlined here have been realized. Most of this report has focused on describing the essential validity elements that partially provide this necessary evidence. The existing evidence was organized into an essential validity framework that can be used to evaluate whether professional testing standards consistent with a Field Test have been met. The essential validity elements presented here constitute critical evidence and are elements that are “relevant to the technical quality of a testing system” (AERA et al., 1999, p. 17). The evidence in support of these essential elements highlighted here referenced the relevant information from the other chapters of this technical report or referenced specific Smarter Balanced supporting documents, the products of other Smarter Balanced workgroups or outside groups. The types of evidence presented here are more consistent with those supporting a Field Test prior to operational administration. Many types of evidence from external sources could not reasonably be collected when so many parts of the program were being developed simultaneously.

The second validity framework consisting of the five sources of validity evidence represents a comprehensive agenda that entails a host of longer-range validation studies. At this juncture, a few potentially important types of validity activities are anticipated. An important area of research is the relationship of Smarter Balanced with other important national and international large-scale assessment programs, such as NAEP, TIMSS, and PISA. This is important to establish the technical properties and rigor of the Smarter Balanced assessments. Most important is the validation of the measurement of college and career readiness, which entails collecting various types of criteria from sources outside the Smarter Balanced assessment system. In considering potential validity studies that will be important in the future, establishing research support systems to enable these activities for outside investigators will have lasting benefits and ones that will augment the validity of Smarter Balanced.

References

- Abedi, J. & Ewers, N. (2013). *Accommodations for English Language Learners and Students with Disabilities: A Research-Based Decision Algorithm*. Smarter Balanced Assessment Consortium.
- Achieve, Inc. (2004). Ready or not: creating a high school diploma that counts. American Diploma Project. Washington, DC: Achieve, Inc. Retrieved February 11, 2008, from www.achieve.org/files/ADPreport_7.pdf.
- ACT (2006). *Ready for college, ready for work. Same or different?* Iowa City, IA: Author.
- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40, 955-959.
- 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.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Institutes for Research (2014a). *Technical Specifications Manual for Online Testing For the Spring 2014 Field Test Administration*. Smarter Balanced Assessment Consortium.
- American Institutes for Research (2014b). Smarter Balanced Scoring Specification: 2014–2015 Administration.
- Betebenner, D. W. (2011). *New directions in student growth: The Colorado growth model*. Paper Presented at the National Conference on Student Assessment, Orlando, FL, June 19, 2011. Retrieved March 29, 2012, from <http://ccsso.confex.com/ccsso/2011/webprogram/Session2199.html>.
- Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22, 21-29.
- Camara, W. J. (2013). Defining and measuring college and career readiness: A validation framework. *Educational Measurement: Issues and Practice*, 32, 16–27.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Castellano, K. E., & Ho, A. D. (2013). *A Practitioner's Guide to Growth Models*. Council of Chief State School Officers.
- Cohen, J. & Albright, L. (2014). *Smarter Balanced Adaptive Item Selection Algorithm Design Report*. American Institutes for Research.
- 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.
- Crocker, L. M., Miller, D., and Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, 2, 179-194.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

- Dorey, N (2014). *Smarter Balanced "Tests of the Test" Successful: Field Test Provides Clear Path Forward*
<http://csai-online.org/resource/698>
- Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span.
Psychological Bulletin, 93, 179-197.
- ETS (2015). *Study of the Relationship Between the Early Assessment Program and the Smarter
Balanced Field Tests*. Prepared for the California Department of Education by Educational
Testing Service.
- ETS (2014). *Online Field Test Administration Manual for Spring 2014 Field Tests of English Language
Arts/Literacy and Mathematics*. Smarter Balanced Assessment Consortium.
- ETS (2002). *ETS Standards for Quality and Fairness*. Educational Testing Service.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence.
Educational Measurement: Issues and Practice, 18, 5-9.
- Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Berk (ed.),
Criterion-referenced measurement: the state of the art. Baltimore: Johns Hopkins University
Press.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of
Educational Research*, 64, 425-461.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting
standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and
perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th ed., pp. 17-64).
Washington, DC: American Council on Education/Praeger.
- Linn, R. L. (2006). The Standards for Educational and Psychological Testing: Guidance in test
development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp.
27-38), Mahwah, NJ: Lawrence Erlbaum.
- Loomis, S. C. (2011, April). *Toward a validity framework for reporting preparedness of 12th graders
for college-level course placement and entry to job training programs*. Paper presented at the
annual meeting of the National Council on Measurement in Education, New Orleans.
- Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and
instruction, *Review of Educational Research* 4, 1332-1361.
- McGraw-Hill Education/CTB (2014). *Smarter Balanced Pilot Automated Scoring Research Studies*.
Smarter Balanced Assessment Consortium.
- McGraw-Hill Education/CTB (2014). *Field Test: Automated Scoring Research Studies*. Smarter
Balanced Assessment Consortium.
- Measured Progress/ETS (2012). *Smarter Balanced Assessment Consortium: Annotated Bibliography:
Item and Task Specifications and Guidelines*. Smarter Balanced Assessment Consortium.
- Measured Progress/ETS (2012). *General Item Specifications*. Smarter Balanced Assessment
Consortium.
- Measured Progress/ETS (2012). *Signing Guidelines*. Smarter Balanced Assessment Consortium.

- Measured Progress/ETS (2012). *Tactile Accessibility Guidelines*. Smarter Balanced Assessment Consortium.
- Measured Progress/ETS (2012). *Performance Tasks Specifications*. Smarter Balanced Assessment Consortium.
- Measurement Incorporated/CTB/McGraw-Hill (2014). *Achievement Level Setting Plan*. Smarter Balanced Assessment Consortium Internal Document.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed.). Washington, DC: American Council on Education.
- Mislevy, R. J. (2009, February). Validity from the perspective of model-based reasoning. *CRESST report 752*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61-90), Mahwah, NJ: Lawrence Erlbaum.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.
- National Center on Educational Outcomes (2015). *Usability, Accessibility, and Accommodations Guidelines*. Smarter Balanced Assessment Consortium.
- National Council on Measurement in Education (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Madison, WI: Author.
- Navigation North Learning (2014). *The Smarter Balanced Technology Strategy Framework and Testing Device Requirements*. Smarter Balanced Assessment Consortium.
- Penfield, R. D., & Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement In Education*, 17, 359-370.
- Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education*, 5, 285-301.
- Porter, A. C., & Smithson, J. L. (2002, April). *Alignment of assessments, standards and instruction using curriculum indicator data*. Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, LA.
- Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). Benchmarking and alignment of standards and testing (Technical Report 566). Washington, DC: Center for the Study of Evaluation.
- Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests*. National Research Council.
- Sireci, S. G. (2012). Smarter *Balanced Assessment Consortium: Comprehensive Research Agenda*. Report Prepared for the Smarter Balanced Assessment Consortium.
- Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299-321.
- Smarter Balanced (2014). *Mathematics Summative Assessment Blueprint*. Smarter Balanced Assessment Consortium.

- Smarter Balanced (2014). *ELA/Literacy Summative Assessment Blueprint*. Smarter Balanced Assessment Consortium.
- Smarter Balanced (2014). *Statements of Support: Achievement Level Setting for the Smarter Balanced Assessments*. Smarter Balanced Assessment Consortium.
- Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Scoring Guide for Selected Short-Text Mathematics Items (Field Test 2014)*.
- Smarter Balanced Assessment Consortium (2014). *Mathematics Performance Task Specifications (Design, Development, and Scoring Plan)*.
- Smarter Balanced Assessment Consortium (2014). *Calculator Availability Information for 2014 Field Test*.
- Smarter Balanced (2013). *Career Readiness Frameworks Introduction and Implementation Guide*.
- Smarter Balanced (2013). *Initial Achievement Level Descriptors and College Content-Readiness Policy*. Smarter Balanced Assessment Consortium.
- Thissen, D. & Wainer, H. (eds.) (2001). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Vasavada, N., Carman, E., Hart, B., & Luisser, D. (2010). Common core state standards alignment: Readiness, PSAT/NMSQT, and SAT. *Research report 2010-5*. New York: The College Board.
- Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*, 7-25.