

Smarter Balanced Assessment Consortium: Comprehensive Research Agenda

Report of Recommendations Prepared by Stephen G. Sireci

December 31, 2012



Acknowledgments

Smarter Balanced is a true collaboration among some of the most talented and dedicated educational researchers in the United States. This research agenda could not have been produced without the help of many of them, particularly Carole Gallagher, Marty McCall, Christyan Mitchell, Joe Willhoft, Joseph Martineau, Vince Dean, Carissa Miller, Steve Slater, Randy Bennett, Jacqueline King, Mohamed Dirir, Liru Zhang, Garron Gianopulos, Patricia Reiss, and April Zenisky. I am also grateful for the valuable input from the Smarter Balanced Technical Advisory Committee and the Validation and Psychometrics/Test Design Work Group. Many of the ideas for validity studies came from conversations with these colleagues.

Table of Contents

I. Introduction	4
Purposes of This Report	4
II. Standards and Guidelines for Test Validation.....	6
The Standards for Educational and Psychological Testing: A Validation Framework	6
NCLB Peer Review Guidelines	8
Other Validation Guidelines	9
III. Smarter Balanced Purpose Statements for Validation.....	11
IV. Essential Validity Elements for Summative and Interim Assessments	16
V. Validity Agenda for Summative Assessments	25
Summative Assessment Purpose 1:.....	25
Summative Assessment Purposes 2 and 3:	35
Validating College and Career Readiness Benchmarks.....	39
Summative Assessment Purpose 4:.....	46
Summative Assessment Purpose 5:.....	49
Summative Assessment Purpose 6:.....	50
Summative Assessment Purpose 7:.....	52
VI. Validity Agenda for Interim Assessments	56
Interim Assessment Purpose 1:.....	56
Interim Assessment Purpose 2:.....	57
Interim Assessment Purpose 3:.....	58
Interim Assessment Purpose 4:.....	59
VII. Research Agenda for Formative Assessment Resources.....	60
VIII. Summary: The Smarter Balanced Assessment Consortium Validity Argument	63
Summarizing the Validity Evidence	64
IX. Ongoing Validation Activities and Support Systems	72
References	73
Appendix A: Smarter Balanced Theory of Action and Derivation of Purpose Statements.....	83
Appendix B: Description of Alignment Methods.....	86
Appendix C: Description of Item Similarity Rating Approach to Evaluating Test Content.....	87
Appendix D: Description of ResidPlots2: IRT Residual Analysis Software.....	89

Smarter Balanced Assessment Consortium Comprehensive Research Agenda

I. Introduction

In September 2010, the U.S. Department of Education awarded \$175 million to the Smarter Balanced Assessment Consortium (Smarter Balanced) to develop assessments in English language arts (ELA) and mathematics that would “provide ongoing feedback to teachers during the course of the school year, measure annual student growth, and move beyond narrowly-focused bubble tests” (U.S. Department of Education, 2010). This award was part of the federal government’s \$4.35 billion Race to the Top competitive grant fund, which rewarded states for:

- Adopting standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy;
- Building data systems that measure student growth and success, and inform teachers and principals about how they can improve instruction;
- Recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most; and
- Turning around our lowest-achieving schools. (U.S. Department of Education, 2009a, p. 2)

The goals of Smarter Balanced are comprehensive and are consistent with those of the Race to the Top Initiative. At the time of this report, Smarter Balanced represents a consortium of 25 states working together to develop cutting-edge ELA and mathematics assessments that feature computer-adaptive technology, technology-enhanced item formats, summative and interim assessments, and formative assessment resources. The assessment system being developed by the Consortium is designed to provide comprehensive information about student achievement that can be used to improve instruction and provide extensive professional development for teachers. The Smarter Balanced assessment system focuses on the need to strongly align curriculum, instruction, and assessment, in a way that provides valuable information to support educational accountability initiatives.

The specific goals of Smarter Balanced are described in its “Theory of Action,” which is presented in Appendix A. The purpose of this report is to outline the research that should be conducted to (a) provide information to Smarter Balanced to help the Consortium accomplish its goals as it implements the program, and (b) evaluate the degree to which the Consortium is meeting its goals. Given that a large part of Smarter Balanced involves developing, administering, and scoring the assessments, and reporting the assessment results, much of the recommended research is based on the guidance provided by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), hereafter referred to as the Standards.

Purposes of This Report

The purposes of this report are to inform Smarter Balanced of research that should be done to evaluate the degree to which the Consortium is accomplishing its goals and to demonstrate that the assessment system adheres to professional and federal guidelines for fair and high-quality assessment. The intent is to provide a comprehensive and detailed research agenda for the Consortium that includes suggestions and guidance for both short- and long-term research activities that will support Consortium goals.

To best inform the Consortium, we provide a description of the Standards, which were used as a framework for developing much of the research agenda. Integral to this description is a discussion of validity and the test validation process. We also reference the U.S. Department of Education's Standards and Assessments Peer Review Guidance (2009b), which stipulated the requirements for assessment programs to receive federal approval under the No Child Left Behind (NCLB) legislation. Although not described in this report, the research agenda also considered and is consistent with the Joint Committee on Standards for Educational Evaluation (JCSEE) Program Evaluation Standards (Yarbrough, Shulha, Hopson, & Caruthers, 2011) as well as the Guiding Principles for Evaluators (American Evaluation Association, 2004), which state that "evaluators aspire to construct and provide the best possible information that might bear on the value of whatever is being evaluated" (p. 1). The research agenda proposed here is designed to provide the best possible information to Smarter Balanced for understanding both the degree to which the Consortium is meeting its goals as well as what it can do to improve the system as it evolves.

In the remainder of this report, we (a) discuss the development of a validation plan that is consistent with the Standards and with the U.S. Department of Education's Standards and Assessments Peer Review Guidance; (b) list the primary purposes and goals of Smarter Balanced; (c) list the key validity issues associated with these purposes and goals; and (d) provide a description of studies that should be done to provide evidence regarding the degree to which Smarter Balanced assessments and activities are meeting the intended goals.

II. Standards and Guidelines for Test Validation

The Standards for Educational and Psychological Testing: A Validation Framework

There have been debates regarding what the term “validity” refers to, but for over 50 years three organizations—the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME)—have worked together to forge a consensus view of validity and provide guidance for developing and validating educational and psychological tests (Sireci, 2009). Currently, the Standards for Educational and Psychological Testing (AERA et al., 1999) define validity as “...the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). This definition emphasizes the importance of theory and empirical evidence to support the use of a test for a particular purpose. Thus, the research agenda for Smarter Balanced must be derived from the intended testing purposes and how assessment scores will be used.

The Standards describe the process of validation as that of developing a 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 Standards. For example,

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. This includes 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 . . . (AERA et al., 1999, p. 17)

This excerpt reinforces the Standards’ emphasis that validation should center on test-score interpretation for specific uses. The research agenda developed for Smarter Balanced will be designed to fulfill the requirements of a sound validity argument as described by the Standards.

The Standards’ Five Sources of Validity Evidence. To develop a sound validity argument, the Standards provide a validation framework based on five sources of validity evidence. These sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing.

Validity evidence based on *test content* refers to traditional forms of content validity evidence such as practice (job) analyses and subject-matter expert review and rating of test specifications and test items (Crocker, Miller, & Franks, 1989; Sireci, 1998), as well as newer “alignment” methods for educational tests that evaluate the links among curriculum frameworks, testing, and instruction (Bhola, Impara, & Buckendahl, 2003; Martone & Sireci, 2009). Evidence in this category is used to confirm that the tests that students take adequately represent the intended knowledge and skill areas. Confirming the degree to which the Smarter Balanced test specifications capture the intended Common Core State Standard (CCSS) and confirming that the items that students take adequately represent the areas delineated in the test specifications are examples of validity evidence based on test content that will be needed to build a strong validity argument for the Smarter Balanced assessments.

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). Such evidence can include interviewing test takers about their responses to test questions, systematic observations of test response behavior, evaluation of the criteria used by judges when scoring performance tasks, analysis of item response time data, and evaluation of the reasoning processes that examinees use when solving test items (Embretson [Whitley], 1983; Messick, 1989; Mislevy, 2009). Such evidence will be needed to confirm that the Smarter Balanced assessments are measuring the cognitive skills that they intend to measure, and that students are using the targeted skills to respond to the test items.

Validity evidence based on *internal structure* refers to statistical analysis of item and sub-score data to investigate the primary and secondary (if any) dimensions measured by an assessment. Procedures for gathering such evidence include factor analysis (both exploratory and confirmatory) and multidimensional scaling. Internal structure evidence also evaluates the “strength” or “salience” of the major dimensions underlying an assessment, and so would also include indices of measurement precision, such as reliability estimates, decision accuracy and consistency estimates, generalizability coefficients, conditional and unconditional standard errors of measurement, and test information functions. In addition, analysis of differential item functioning (DIF), which is a preliminary statistical analysis to assess item bias, also falls under the internal structure category.

Evidence based on *relations to other variables* refers to traditional forms of criterion-related validity evidence, such as concurrent and predictive validity studies, as well as more comprehensive investigations of the relationships among test scores and other variables, such as multitrait-multimethod studies (Campbell & Fiske, 1959), and score differences across different groups of students, such as those who have taken different courses. 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), to evaluate 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 will be essential for supporting the validity of certain inferences based on scores from Smarter Balanced assessments (e.g., certifying college and career readiness).

Finally, evidence based on *consequences of testing* refers to 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 and job applications. Other investigations of testing consequences relevant to the Smarter Balanced goals include analysis of students’ opportunity to learn the CCSS, and analysis of changes in textbooks and classroom artifacts. With respect to educational tests, the Standards stress studying testing 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).

Thus, it is important that validity evidence based on testing consequences is prominent in the Smarter Balanced research agenda.

Using the Standards as a Validation Framework. 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). Therefore, they have great utility in

guiding a validity agenda. The validation research component of this comprehensive research agenda is based on crossing the intended purposes and use of Smarter Balanced assessments with the Standards' five sources of validity evidence. Therefore, the first step in determining the Smarter Balanced validity research agenda was to explicitly state its goals and purposes. These goals and purposes that are the focus of validation are described in Chapter III of this report.

NCLB Peer Review Guidelines

One of the seven principles underlying the Smarter Balanced Theory of Action is the adherence “to established professional standards” (Smarter Balanced, 2010, p. 33). In addition to adhering to the Standards, the Consortium will also meet the requirements of the U.S. Department of Education’s Peer Review process for NCLB assessments. Although these requirements are temporarily suspended as they undergo revision (Delisle, 2012), they remain important because they reflect the Department’s most recent standards for ensuring quality and equity in statewide assessment programs. Thus, the research agenda incorporates much of the guidance provided in the Standards and Assessments Peer Review Guidance (U.S. Department of Education, 2009b). There is a great deal of overlap between the Standards and the U.S. Department of Education’s Peer Review Guidance. However, the Guidance stipulates several important requirements that are highlighted in this research agenda. In particular, it requires:

- Providing evidence of the purpose of an assessment system and studies that support the validity of using results from the assessment system for their stated purpose and use (p. 42)
- Strong correlations of test and item scores with relevant measures of academic achievement, and weak correlations with irrelevant characteristics, such as demographics (p. 42)
- Investigations regarding whether the assessments produce intended or unintended consequences (p. 42)
- Documentation supporting evidence of the delineation of cut scores and the rationale and procedures for setting cut scores (pp. 21–22)
- Evidence of the precision of the cut scores & consistency of student classification (p. 44)
- Evidence of reliability for overall population and for each reported subpopulation (p. 44)
- Evidence of alignment over time through quality control reviews (p. 52)
- Evidence of comprehensive alignment and measurement of the full range of content standards and depth of knowledge and cognitive complexity (p. 54)
- Evidence that the assessment plan and test specifications describe how all content standards are assessed and how the domain is sampled to lead to valid inferences about student performance on the standards, individually and in the aggregate (using impartial experts in the process) (p. 54)
- Scores that reflect the full range of achievement standards (p. 57)
- Documentation to describe that the assessments are a “coherent” system across grades and subjects including studies establishing vertical scales (p. 34)
- Identification of how each assessment will provide information on the progress of students (p. 34)

The overlap of these requirements with the Standards is clear, and the anticipated revisions to this guidance will likely retain these key features. For example, in the recent letter informing states of the temporary suspension of peer review, the Department reiterated the following desired characteristics:

A high-quality assessment system [is] one that is “valid, reliable, and fair for its intended purposes; and measures student knowledge and skills against college- and career-ready standards in a way that

- Covers the full range of those standards, including standards against which student achievement has traditionally been difficult to measure;
- As appropriate, elicits complex student demonstrations or applications of knowledge and skills;
- Provides an accurate measure of student achievement across the full performance continuum, including for high- and low-achieving students;
- Provides an accurate measure of student growth over a full academic year or course; produces student achievement data and student growth data that can be used to determine whether individual students are college- and career-ready or on track to being college- and career-ready;
- Assesses all students, including English language learners and students with disabilities;
- Provides for alternate assessments based on grade-level academic achievement standards or alternate assessments based on alternate academic achievement standards for students with the most significant cognitive disabilities, consistent with 34 C.F.R. § 200.6(a)(2); and
- Produces data, including student achievement data and student growth data, that can be used to inform: determinations of school effectiveness for purposes of accountability under Title I; determinations of individual principal and teacher effectiveness for purposes of evaluation; determinations of principal and teacher professional development and support needs; and teaching, learning, and program improvement.”

These characteristics of high-quality assessment systems were also considered in development of the comprehensive research agenda to ensure that evidence will be provided to demonstrate that the Smarter Balanced system meets these high standards.

Other Validation Guidelines

In addition to the AERA et al. (1999) Standards and the U.S. Department of Education’s (2009) Peer Review Guidance, there have been other seminal works that have influenced test validation practices. Messick’s (1989) landmark chapter influenced the Standards and encouraged validators to focus on test use and the evaluation of testing consequences. Kane (1992, 2006), mentioned earlier, advanced Cronbach’s (1988) notion of validation as an evaluation argument, and this notion is also embodied in the Standards. A recent addition to the validity literature is Bennett (2010), who expanded discussion of validation to include validation of a theory of action. This perspective is relevant to Smarter Balanced and is addressed in Chapter VIII. In short, this comprehensive research agenda incorporates many of the current theories and practices in test validation.

In addition to general guidelines on validation, there are also guidelines for specific testing applications. For example, the International Test Commission (ITC) produced Guidelines for

Translating and Adapting Tests (Hambleton, 2005; ITC, 2010), which are relevant to the evaluation of the Spanish-language versions of the Smarter Balanced mathematics assessments. There are also guidelines for universal test design (e.g., Johnstone, Altman, & Thurlow, 2006), and sensitivity review (e.g., Ramsey, 1993), which are relevant to the evaluation of the development of the Smarter Balanced assessments. Other documents consulted to guide this research agenda include Kane's (1994, 2001) criteria for evaluating standard setting studies (described further in Chapter IV) and the recent guidelines published by NCME (2012) on maintaining test integrity .

III. Smarter Balanced Purpose Statements for Validation

As mentioned earlier, *validation* refers to gathering and evaluating evidence with respect to specific testing purposes. Thus, a first step in developing the comprehensive research agenda was identifying and articulating the intended purposes of Smarter Balanced. As the AERA et al. (1999) Standards state, “When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described . . .” (p. 168).

Although the Smarter Balanced Theory of Action described the overall goals of the Consortium, it was too general for evaluation or validation purposes. Thus, several steps were conducted to articulate the primary purposes and goals of Smarter Balanced that would be the focus of validation. These steps involved:

1. Extensive review of Smarter Balanced documentation;
2. Compiling a list of explicit claims, goals, and purposes;
3. Presenting this list to the Smarter Balanced Technical Advisory Committee (TAC);
4. Refining the list based on feedback;
5. Presenting the revised list to Smarter Balanced work groups;
6. Observing the Smarter Balanced Collaboration Conference and discussing goals, purposes, and validation plans with work groups, staff, and contractors;
7. Developing a draft list of Smarter Balanced goals and purposes to be the focus of validation;
8. Discussing this list with Smarter Balanced work groups via WebEx teleconferences; and
9. Revising the list based on work group input.

The identification of Smarter Balanced-specific goals began with the Theory of Action (Appendix A), but also involved a review of numerous Smarter Balanced documents, including the original Race to the Top application (Smarter Balanced, 2010), test specification documents (e.g., ETS, 2012a, 2012b), press releases, and requests for proposals (RFPs). More than 50 documents were reviewed in order to detect any stated claims, purposes, or goals. These reviews led to a preliminary list of goals and purposes that were presented to the Smarter Balanced TAC in July 2012. Feedback was received from the TAC and then from selected members of the Smarter Balanced Validation and Psychometrics/Test Design Work Group. Based on this feedback, refinements were made to the list of goals and purposes and were shared with Smarter Balanced leadership at the Collaboration Conference in September 2012. Further feedback was received, which included receipt of other documents that should be factored into the final articulation of goals and purposes.

Based on the observations and interaction with Consortium members, and the feedback provided by the TAC and the work group, a focus-group protocol was developed to involve Smarter Balanced leadership in the final articulation of testing purposes via WebEx teleconferences. Focus groups were held via WebEx in October 2012 with both the Validation and Psychometrics/Test Design Work Group and the Test Administration/Student Access Work Group. Excluding the facilitator, ten people participated in the first focus group (October 24, 2012) and sixteen people participated in the second (October 31, 2012). Each focus group was 90 minutes in duration. Following each focus group, draft purpose statements were sent to the participants via SurveyMonkey, and participants rated and commented on the appropriateness of the draft purpose statements. Based on these ratings and comments, the draft statements

were revised. These statements were presented to the TAC on December 12, 2012, and additional feedback was received and incorporated.

The final list of Smarter Balanced purpose statements that are the focus of validation follow. A description of the Smarter Balanced Theory of Action is presented in Appendix A to illustrate the degree to which the final list of purpose statements covers the major intentions stated in the Theory of Action.

The Smarter Balanced purpose statements for validation are separated into three categories that refer to (a) the summative assessments, (b) the interim assessments, and (c) formative assessment resources.

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

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

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 measured in ELA and mathematics by the summative assessments.
2. Students' performance at the content cluster level, so that 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 and mathematics, to determine whether teaching and learning are on target.
4. Student progress toward the mastery of skills measured in ELA and mathematics *across all students and subgroups of students*.

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. Help teachers and other educators use 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 remainder of this report centers on these purpose statements and their validation. The validation framework for the summative and interim assessments is based on the aforementioned five sources of validity evidence described in the Standards and involves crossing the purpose statements with each of the five sources. The formative assessment resources are not assessments per se, and so the research in support of their intended purposes extends beyond the five sources of validity evidence and follows a more traditional program evaluation approach.

As a prelude to Chapters V and VI, Tables 1 and 2 illustrate the validation framework for the Summative and Interim Assessments by crossing 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 is most important for validating each specific purpose. This presentation is extremely general, but indicates the comprehensiveness of the research agenda. It is also useful for understanding which sources of validity evidence are most important to specific purposes. For example, for purposes related to providing information about students' knowledge and skills, validity evidence based on test content will always be 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, because that evidence includes information regarding decision consistency and accuracy.

Table 1. Validity Framework for Smarter Balanced *Summative Assessments*

The purposes of the Smarter Balanced <i>summative</i> assessments are to provide valid, reliable, and fair information about:	Source of Validity Evidence				
	Content	Internal Structure	Relations w/ Ext. Variables	Response Processes	Testing Consequences
1. Students' ELA and mathematics achievement with respect to those CCSS measured by the ELA and mathematics summative assessments.	√	√	√	√	
2. Whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA and mathematics to be on track for achieving college readiness.	√	√	√		√
3. Whether grade 11 students have sufficient academic proficiency in ELA and mathematics to be ready to take credit-bearing college courses.	√	√	√		√
4. Students' annual progress toward college and career readiness in ELA and mathematics.	√	√	√		√
5. How instruction can be improved at the classroom, school, district, and state levels.	√				√
6. Students' ELA and mathematics proficiencies for federal accountability purposes and potentially for state and local accountability systems.	√	√	√		√
7. Students' achievement in ELA and mathematics that is equitable for <i>all students and subgroups of students</i> .	√	√	√	√	√

Table 2. Validity Framework for Smarter Balanced *Interim Assessments*

The purposes of the Smarter Balanced <i>interim</i> assessments are to provide valid, reliable, and fair information about:	Source of Validity Evidence				
	Content	Internal Structure	Relations w/ Ext. Variables	Response Processes	Testing Consequences
1. Student progress toward mastery of the skills measured in ELA and mathematics by the summative assessments.	√	√		√	
2. Students' performance at the content cluster level, so that 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 and mathematics, to determine whether teaching and learning are on target.		√	√		√
4. Student progress toward the mastery of skills measured in ELA and mathematics <i>across all students and subgroups of students</i> .	√	√	√	√	√

IV. Essential Validity Elements for Summative and Interim Assessments

Before describing specific studies associated with each of the testing purposes listed in the previous chapter, it is important to first consider the fundamental validity information that is needed for *any* educational assessment program. These “essential elements” cut across the five sources of validity evidence and so deserve particular attention. The Standards describe such fundamental information 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” (AERA et al., 1999, p. 17). Most of these essential elements fall under the categories of validity evidence based on test content (e.g., careful test construction) and internal structure (adequate score reliability, scaling, equating), but others, such as test administration and scoring, and careful attention to fairness, fall outside these two categories and do not neatly fit into the others. In addition to these fundamental elements, two other elements are essential: (a) equitable participation and access, and (b) test security.

In this chapter, we describe the types of information needed to confirm that these essential elements are adequately addressed in the research agenda. Because these elements refer to assessments, they are described in relation to the summative and interim assessments. However, “equal participation and access” is also important with respect to the formative assessment resources, which are discussed in Chapter VII.

In Table 3, we present a brief description of the validity evidence for the essential elements associated with the summative and interim assessments. Although the preceding quote from the Standards mentions adequate “reliability,” we refer more generally to adequate “measurement precision” to underscore the need for measurement error to also be conceptualized in other frameworks such as item response theory (IRT) and generalizability theory.

The types of evidence listed in Table 3 will resurface when considering validity evidence for the specific purposes described earlier. This reoccurrence underscores the fundamental nature of these elements for supporting the use of Smarter Balanced assessments for their intended purposes. Most of these essential elements are typically addressed in technical manuals that support an assessment program. Descriptions of the types of studies to be conducted for each essential element follow.

Careful Test Construction

As indicated in Table 3, validity evidence of careful test construction can come from a comprehensive audit of the test development process. This audit should be a comprehensive review of all test development activities, starting with the descriptions of testing purposes, operational definitions of the constructs measured, item development, content reviews, alignment studies, sensitivity reviews, pilot testing, item analyses, DIF analyses, item selection, item calibration, scoring rubrics for constructed-response items, and creation of test booklets (and clarity of test instruction). For adaptive assessments, the adequacy of the item selection algorithm, and the stopping rule, should also be reviewed.

Table 3. Validity Evidence Associated with Essential Elements for Summative and Interim Assessments

Essential Element	Validation Evidence
Careful Test Construction	Audit 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 calibration
Adequate Measurement Precision	Analysis of test information, conditional standard errors of measurement, decision accuracy, decision consistency, and reliability estimates for all reported scores
Appropriate Test Administration	Audit of test administration procedures, analysis of test irregularities, analysis of use and appropriate assignment of test accommodations
Appropriate Scoring	Audit of scoring procedures (hand, automated), inter-rater reliability analyses, rater drift (scale stability) analyses, computer/human comparisons (if relevant), generalizability studies, fairness for minorities
Accurate Scaling and Equating	Third-party verification of horizontal and vertical equating, IRT residual analysis, analysis of equating error, documentation of scaling and equating procedures, population invariance of equating
Appropriate Standard Setting	Comprehensive standard setting documentation, including procedural, internal, and external validity evidence for all achievement level standards set on assessments; includes criterion-related studies
Careful Attention to Fairness	Sensitivity review, DIF analyses, differential predictive validity analyses, comparability analyses (for language and disability accommodations), review of accommodation policies, implementation of accommodations, qualitative and statistical analyses of accommodated tests
Equitable Participation and Access	Analysis of participation rates, test accommodations, translations, and other policies
Adequate Test Security	Analysis of data integrity policies, test security procedures, monitoring of test administrations, analysis of cheating behavior, analysis of item exposure, review of chat rooms and websites for exposed items, review of anomalous results

Examples of types of evidence that would be reviewed are presented in Table 4. Although a checklist format is used in Table 4, an audit would not simply check whether the activity was in place; rather, it would evaluate the quality of the activity.

Table 4. Sample Checklist for Audit of Test Construction Procedures

Activity	Completed	Not Completed	Comments
Theory of Action/testing purposes clearly stated			
Development of test specifications sufficiently documented			
Item writers appropriately trained or recruited			
Items adhere to item writing guidelines			
Items reviewed for content quality and technical adequacy			
Content validity/alignment studies			
Sensitivity reviews			
Pilot study is adequate and representative			
Item analysis (classical)			
DIF analysis			
Item selection based on statistical and content criteria			
Item calibration			
Scoring rubrics for constructed-response items reviewed			
Adaptive item selection algorithm documented			
Test booklets are error-free			

Adequate Measurement Precision

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 classification based on a test score. Examples of this information include estimates of score reliability, standard errors of measurement, conditional standard errors of measurement, item and test information functions, conditional standard error functions, and estimates of decision accuracy and consistency. Estimates of score reliability include internal consistency estimates based on a single test administration (coefficient alpha, stratified alpha, marginal reliability), and those based on testing individuals more than once (test retest, parallel forms). The essential information needed for the Smarter Balanced assessments includes reliability estimates for all scores reported for students, estimates of decision consistency and accuracy for any reported achievement level results, and the traditional test information and standard error functions associated with IRT analyses. Generalizability studies that focus on specific sources of error will be important for identifying the sources of measurement error.

Appropriate Test Administration

Evidence in this category involves review of test administration manuals and other aspects of the test administration processes. This review should 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, should also be included. The policies and procedures for granting and providing accommodations to students with disabilities and English language learners should also 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.

Appropriate Scoring

Validity evidence to confirm that the scoring of Smarter Balanced assessments is appropriate should include a review of scoring documentation. The Standards state that such documentation “should be presented . . . in sufficient detail and clarity to maximize the accuracy of scoring” (AERA et al., 1999, p. 47), as should the processes for selecting, training, and qualifying scorers. The scoring processes should also include monitoring of the frequency of scoring errors and how they are corrected. In terms of specific studies, evaluation of scorer reliability and score scale drift should be conducted. If any assessments are scored locally, the degree to which the scorers are trained, and the accuracy of their scores, should also be studied. Evidence in this category should also confirm that the routing of students during the adaptive exams is correct, and that all computerized scoring programs are accurate. The Standards *also* point out that one way to evaluate computerized scoring algorithms is to commission “an independent review of the algorithms by qualified professionals” (p. 70). Generalizability studies to locate sources of measurement error due to scoring will also provide important evidence.

Accurate Scaling and Equating

Scaling and equating are essential activities for providing valid scores and score interpretations for Smarter Balanced assessments. Scaling activities include item calibration and creation of the standardized scale on which scores are reported. Equating activities will ensure that different forms of the assessments are on a common scale, as are scores reported over time. At the time of this writing, the summative assessments are intended to be vertically equated across grades. For the adaptive tests, the notion of a test “form” does not apply because the items are calibrated onto a common scale and can be assembled together uniquely for each examinee. This process requires that the items are correctly calibrated and that the IRT model sufficiently fits the data. Validity evidence for scaling and equating will include evaluation of the IRT model, confirming the hypothesized dimensionality of the assessments, evaluating equating documentation and estimates of equating error, evaluating the viability of a single construct (dimension) across grades, and, potentially, evaluating the invariance of the equating functions across important subgroups of students, such as students in different states. If funds are available, a “redundancy analysis,” where an independent third party replicates the equating done by the contractor, would provide an important validity check on the accuracy of the equating.

Appropriate Standard Setting

When achievement level standards are set on tests, test scores often become less important than the classifications that students receive. The standard setting literature is full of 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 will use achievement levels, some of which will signify that students are “on track” to college readiness (grades 3–8) or “college ready” (grade 11). Kane (1994, 2001) wrote about gathering and documenting validity evidence for standards set on educational tests and categorized the evidence into three categories—procedural, internal, and external.

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 were 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 cutscores 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, 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 the students' test-based achievement level classifications. It is likely that external validity evidence will be particularly important for validating the "college and career readiness" standards set on the summative assessments because several measures of college readiness already exist. In addition to classification consistency, the degree to which the constructs measured by these assessments overlap with the Smarter Balanced summative assessments, and the degree to which their definitions of readiness are similar, should be studied.

Some specific criteria that can be used to provide validity evidence for standard setting are summarized in Table 5. This table, adapted from Sireci, Hauger, Wells, Shea, & Zenisky (2009), illustrates the activities that should be conducted to (a) facilitate validity within the standard setting study, (b) evaluate the validity of the standard setting after it has been completed, or (c) do both.

Table 5. Summary of Criteria for Evaluating Standard Setting Studies

Evidence	Criterion	Brief Explanation
Procedural	Care in selecting participants	Qualifications, competence, and representativeness of panelists; sufficient number of panelists
	Justification of standard setting method(s)	Degree to which methods used are logical, defensible, and congruent with testing purpose
	Panelist training	Degree to which panelists were properly oriented, prepared, and trained
	Clarity of goals/tasks	Degree to which standard setting purposes, goals, and tasks were clearly articulated

Evidence	Criterion	Brief Explanation
	Appropriate data collection	Data were gathered as intended
	Proper implementation	Method was implemented as intended
	Panelist confidence	Panelists understood tasks and had confidence in their ratings
	Sufficient documentation	Documentation of the entire process so that (a) it is understood and (b) it can be replicated
Internal	Sufficient inter-panelist consistency	Reasonable standard deviations and ranges of cut scores across panelists
	Decreasing variability across rounds	The variability across panelists' cut scores decreases across rounds—evidence of emerging consensus
	Small standard error of cut score (consistency within method)	Estimate of degree to which cut scores would change if study were replicated
	Consistency across independent panels	Estimate of degree to which cut scores would change if different panelists were used
	Consistency across panelist subgroups	Estimate of degree to which cut scores would change if specific types of panelists were used
	Consistency across item formats	Estimate of the consistency of cut scores across item formats (e.g., SR, CR items)
	Analysis of borderline students' performance on specific items	Degree to which expectations of hypothetical borderline students' performance are consistent with the performance of students near the cut scores
External	Consistency across standard setting methods	Degree to which results from different standard setting methods yield similar results
	Consistency across other student classification data	Degree to which classifications of students based on external data are congruent with classifications based on the cut scores
	Mean differences across proficiency groups on external criteria	Degree to which students classified into different achievement levels differ on other relevant variables
	Reasonableness	Degree to which cut scores produce results that are within a sensible range of expectations

Note: Adapted from Sireci et al. (2009).

Careful Attention to Fairness

Careful attention to fairness begins at the earliest stages of test development and includes many of the activities described in the previous section on careful test construction. One important aspect of

fairness is acknowledging the diversity within the student population when defining the constructs measured. Considerations of this diversity will reduce ethnocentricity in the construct definition and allow the development of accommodations policies that stay faithful to the construct measured. Sensitivity reviews and analysis of DIF and differential predictive validity are other important aspects of test fairness. Ensuring that students have the opportunity to learn material before it is tested and ensuring that a fair appeal process is in place are other important aspects of fairness. The presence of these practices and policies will be checked as part of the research agenda. The recent NCME document on data integrity underscores the need for testing programs to have policies and procedures to “ensure that all students have appropriate, fair, and equal opportunities to show their knowledge, skills, and abilities” (NCME, 2012, p. 3).

Equitable Participation and Access

The Smarter Balanced system is designed for *all* students, and the intent is to provide flexibility and remove barriers that may inhibit students from taking the test and performing their best. The system is also designed to provide information widely, in transparent fashion, to all stakeholders. Equitable participation and access ensures that all students can take the test in a way that allows them to comprehend and respond appropriately.¹ The research agenda should include an analysis of participation rates across subgroups of students as well as a review of the procedures in place to ensure full participation. In particular, the degree to which Smarter Balanced offers sensible accommodations for students with disabilities and English language learners should be studied, as well as the availability and successful implementation of those accommodations. As stated in the recent NCME (2012) guidelines on test integrity, “Students who need accommodations due to language differences or students with disabilities may require appropriate modifications to materials and administrative procedures to ensure fair access to the assessment of their skills” (p. 3).

The U.S. Department of Education’s Peer Review Guidance (2009b) provides additional guidance for confirming equitable participation and access. For example, it requires:

- Evidence of judgmental and data-based steps to ensure that assessments are fair and accessible to all students (p. 45)
- Evidence of how universal design or linguistic accommodations are incorporated (p. 45)
- Evidence that students with disabilities were included in the development process (p. 45)
- A policy on appropriate selection and use of accommodations (p. 47)
- Routine monitoring of accommodations used and ensuring that those used are used during instruction (p. 49)
- Checks of quality and consistency for accommodations given to English language learners (p. 49)
- Analysis of effect of usage of accommodations for English language learner students and students with 504s and IEPs (p. 49)

Another aspect of equitable participation and access is the provision of opportunities to retake an assessment. According to current policy, Smarter Balanced “will offer a retake opportunity on the CAT portion of the summative assessment for students who feel their scores are inaccurate or that believe the test was administered under non-standard circumstances” (Smarter Balanced, n.d.).

¹ Marty McCall, personal communication, December 22, 2012.

Adequate Test Security

Test security is a prerequisite to validity. Threats to test security include cheating behaviors by students, teachers, or others who have 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 of students' answers to test questions. All of these behaviors have been observed in the past, and so those who value the validity of test scores worry about the prevalence of cheating behaviors. 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."

Thankfully, there are many proactive steps that testing agencies can take 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 recent 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. Activities during testing include adequate proctoring to prevent cheating, imposters, and other threats. After testing, forensic analysis of students' responses and answer changes, and of 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 for the secure Smarter Balanced assessments will involve a review of the test security procedures and data forensics. The NCME (2012) document on test data integrity should be used to guide this evaluation. This document suggests that security policies should address:

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 respect to specific studies that could evaluate security, in addition to an audit of test security policies, regular and systematic study of incorrect answer patterns for students who took the test in the same setting may be useful. However, with adaptive assessments, the probability of students receiving the same items at similar times is very low. Analyses of large score changes over time may be more useful, but it is important that any students, classes, or schools flagged for large score gains be considered innocent until proven guilty using external data (Wainer, 2011, chapter 8). Finally, given that most Smarter Balanced assessments will be delivered via computer, analysis of the time that students take to respond to items (e.g., are they correctly answering items in less time than it takes to read the item), and when tests are being accessed (are some tests accessed after hours?) will also provide important information regarding test security. Appendix C of the NCME (2012) document lists other examples of forensic analyses that could be conducted to evaluate test security.

Summary of Essential Validity Elements

In considering the essential validity elements that are "relevant to the technical quality of a testing system" (AERA et al., 1999, p. 17), we arrive at many of the studies that should be contained within

the comprehensive research agenda. These studies will be highlighted again in the remaining chapters to underscore how they provide important information relevant to specific purposes of the Smarter Balanced Assessment Consortium, and are coordinated with the other studies described in the Introduction to this report.

V. Validity Agenda for Summative Assessments

As described in Chapter III, there are seven purposes associated with the Smarter Balanced Summative Assessments that we recommend be the focus of validation. All of the studies discussed in Chapter IV that pertain to essential validity elements apply to these purposes. In this chapter, we relate these studies to each purpose statement and provide further descriptions where necessary.

It is important to note that each of the summative assessment purpose statements in Chapter III has the common preface “The purposes of the Smarter Balanced *summative* assessments are to provide valid, reliable, and fair information about . . .” In the sections that follow, we specify each purpose statement and then discuss the studies that should be done to provide the evidence to support the validity of the purpose. Within each purpose, the studies are organized by the Standards’ five sources of validity evidence.

Summative Assessment Purpose 1:

Provide valid, reliable, and fair information about students’ ELA and mathematics achievement with respect to those CCSS measured by the ELA and mathematics summative assessments.

As indicated in Table 1 (p. 14), validity evidence to support this purpose should come from at least three sources—test content, internal structure, and response processes. With respect to validity evidence based on test content, studies should be conducted to confirm that the content of the summative assessments adequately represents the CCSS intended to be measured in each grade and subject area. Appraisals of content domain representation and congruence to the CCSS must be made by carefully trained and *independent* subject-matter experts, not by employees of or consultants for the testing contractors. Validity evidence based on internal structure should involve analysis of item response data to confirm that the dimensionality of those data match the intended structure and support the scores that are reported. All 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 types of studies that are recommended for each of these three sources of validity evidence are described next.

Validity Studies Based on Test Content. Validity studies based on test content for the Smarter Balanced summative assessments need to evaluate the degree to which the assessments adequately measure the CCSS that they are designed to measure and in a way that conforms to the intended *evidence-centered design* (ECD; Mislevy & Riconscente, 2006). There should be at least two levels to the analysis. The first level would evaluate the degree to which the test specifications for the assessment sufficiently represent the intended CCSS. The second level of analysis should evaluate the degree to which the items administered to students adequately represent the test specifications. Studies relevant to these levels include traditional content validity studies (e.g., Crocker et al., 1989) and alignment studies (Bhola et al., 2003; Martone & Sireci, 2009; Porter & Smithson, 2002; Rothman, 2003; Webb, 2007). In Appendix B, we present brief descriptions of traditional content validity and alignment approaches and how they relate to one another.

Evaluating test specifications. To evaluate the appropriateness of the test specifications, the process by which the specifications were developed should be reviewed to ensure that all member states had input and that there was consensus regarding the degree to which the test specifications represent the CCSS targeted for the assessment. The degree to which states agree that the test specifications appropriately represent the CCSS, given the constraints of the assessment, could be ascertained by surveying curriculum specialists in the departments of education in the member states. Surveys could be constructed where these specialists would respond to selected- and open-response questions that would require them to comment on the degree to which the test specifications

adequately define the CCSS intended to be measured on the summative assessments, and the degree to which the relative weights of the cells in the test specifications reflect the corresponding emphases in the CCSS.

Evaluating content and cognitive representation. To evaluate the degree to which the summative assessments adequately represent the test specifications requires recruiting and training qualified and independent *subject-matter experts* (SMEs) in ELA, writing, and mathematics to review the CCSS within the test specifications and Smarter Balanced test items. At least two hypothesized aspects of the assessments need to be validated using SMEs. First is that the items are appropriately measuring the CCSS that they are designed to measure. Second is that the items are measuring the breadth of higher- and lower-order cognitive skills that they are designed to measure. There are a variety of methods that could be used to evaluate these aspects of content validity—some based on traditional notions of content validity, and others based on alignment methodology (Martone & Sireci, 2009). What the specific method is called is not important. What is important is that the tasks presented to the SMEs allow them to provide the data needed to evaluate the degree to which the assessments sufficiently represent the intended CCSS and the cognitive skills targeted by these standards.

To evaluate the degree to which each test item adequately represents (i.e., is aligned with) its corresponding CCSS, there are several studies that could be conducted, ranging from simply having SMEs match test items to claim areas (similar to Webb’s categorical concurrence or Achieve’s [2006] blueprint confirmation) to having the SMEs use a Likert-type rating scale to rate the congruence between each item and the CCSS that it is designed to measure. An example of the “matching” approach is presented in Figure 1, and an example of how the data from such a study could be summarized is presented in Figure 2. An example of the rating approach is presented in Figure 3; an example of how the rating scale data can be summarized is presented in Figure 4.

Regardless of the method chosen, appropriately summarizing the results of these content-based validity studies is important. Results should be analyzed at the item level to screen out or revise any items that have poor alignment ratings. More important, however, is aggregating the data so that the representation of the claims or assessment targets within each subject area can be evaluated.

In addition to the descriptive summaries of alignment, these studies should also compute congruence/alignment statistics. Such statistical summaries range from purely descriptive to those that involve statistical tests. On the descriptive end, Popham (1992) suggested a criterion of 7 of 10 SMEs rating an item congruent with its standard to confirm the fit of an item to its standard. This 70% criterion could be applied to the claim level and other aggregations of items. On the statistical end, 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 SMEs’ mean ratings of content congruence.

Figure 1. Sample Item/Assessment Target Rating Form for Summative Assessment: Reading (Literary)

Item #	Assessment Target (choose one for each item)						
	Key Details	Central Ideas	Word Meanings	Reasoning & Evaluation	Analysis w/in, across Texts	Text Structures & Features	Language Use
432							
433							
434							
443							
563							
578							
579							
580							
581							

From the matching approach (Figure 1), we can see how these data can inform us about the degree to which the assessment targets are represented by the items in a general sense. For example, in Figure 2, we see that the items associated with the assessment target “Analysis within and across Texts” were generally considered congruent with this target by the SMEs, but the items measuring “Language Use” were less congruent. Specific items could be revised or deleted to improve the representation of an assessment target. However, the matching approach does not give us information about *how well* the items measure their associated achievement target. Therefore, the rating scale approach is preferable, even though it may take slightly longer for the SMEs to provide those ratings.

Figure 2. Example Summary of Item/Assessment Target Congruence

Assessment Target	# of Items	% of Items Classified Correctly by All SMEs	% of Items Classified Correctly by at Least 7 SMEs
Key Details	22	45%	86%
Central Ideas	17	88%	94%
Word Meanings	33	55%	97%
Reasoning & Evaluation	25	48%	80%
Analysis w/in, across Texts	12	92%	100%
Text Structures & Features	21	71%	90%
Language Use	17	41%	76%
Average:		56%	89%

Using the rating scale approach (Figure 3), we can get an idea of how well specific items, and the group of items comprising a content category or other level of the test specifications, adequately measure the intended standard or area, with respect to the characteristics of the rating scale. For example, the fictitious results in Figure 4 may suggest that the content categories have good representation with respect to the degree to which the items are measuring the CCSS within each

area. However, some specific items should be flagged for review and possibly revised or deleted. A similar rating task could be used to evaluate how well the items are measuring the intended cognitive skills. A cognitive skill dimension was not noted in the current test blueprints for the Smarter Balanced summative assessments, and so a cognitive skill classification such as that used in the Webb (1999), Achieve (2006), or Porter & Smithson (2002) alignment approaches could be adopted and arranged as a rating task, such as those presented in Figure 1 and Figure 3.

Figure 3. Example of SME Rating Task Assessing Item/CCSS Congruence

Directions: Please read each item and its associated benchmark. Rate how well the item measures its benchmark, using the rating scale provided. Be sure to circle one rating for each item.

Item	Common Core State Standard (Grade 4 ELA)	How well does the item measure its CCSS? (circle one)						Comments (Optional)
		1 (Not at all)	2	3	4	5	6 (Very well)	
226	Refer to details and examples in a text when explaining what the text says explicitly and when drawing inferences from the text.	1	2	3	4	5	6	
238	Determine a theme of a story, drama, or poem from details in the text; summarize the text.	1	2	3	4	5	6	
1006	Describe in depth a character, setting, or event in a story or drama, drawing on specific details in the text (e.g., a character's thoughts, words, or actions).	1	2	3	4	5	6	
1064	Determine the meaning of words and phrases as they are used in a text, including those that allude to significant characters found in mythology (e.g., Herculean).	1	2	3	4	5	6	
1428	Explain major differences between poems, drama, and prose, and refer to the structural elements of poems (e.g., verse, rhythm, meter) and drama (e.g., casts of characters, settings, descriptions, dialogue, stage directions) when writing or speaking about a text.	1	2	3	4	5	6	
1614	Determine a theme of a story, drama, or poem from details in the text; summarize the text.	1	2	3	4	5	6	
1658	Determine the meaning of words and phrases as they are used in a text, including those that allude to significant characters found in mythology (e.g., Herculean).	1	2	3	4	5	6	
1676	Compare and contrast the point of view from which different stories are narrated, including the difference between first- and third-person narrations.	1	2	3	4	5	6	
1733	Refer to details and examples in a text when explaining what the text says explicitly and when drawing inferences from the text.	1	2	3	4	5	6	

Figure 4. Example Summary of Results from Item/CCSS Congruence Study

Item	Content Category	Mean	Median	Aiken Index
226	Reading-Literary	4.2	4.0	.89*
238	Reading-Literary	5.3	5.0	.91*
1006	Reading-Literary	4.1	4.5	.90*
1064	Reading-Literary	3.5	4.0	.91*
1121	Reading-Literary	4.6	4.0	.93*
1214	Reading-Literary	3.7	4.0	.92*
1876	Reading-Literary	5.2	5.0	.95*
	Average for Category	4.4	4.4	.92
1614	Reading-Informational	3.4	3.5	.76*
1658	Reading-Informational	4.5	5.0	.90*
1676	Reading-Informational	5.6	5.5	.95*
1733	Reading-Informational	5.2	5.0	.92*
1963	Reading-Informational	5.4	5.5	.94*
1980	Reading-Informational	5.3	5.5	.93*
1992	Average for Category	4.9	5.0	.90

Notes: Statistics based on 10 SMEs and rating scale where 1 = Not at all, 6 = Very well. * $p < .05$.

Given that data from the rating approach can be aggregated and summarized for each of the dimensions comprising the test blueprints, we recommend this approach, which can be implemented by having SMEs review each item and rate the degree to which it appropriately measures the CCSS it is designed to measure. Based on the literature (e.g., O'Neil, Sireci, & Huff, 2004; Penfield & Miller, 2004), we recommend that at least 10 SMEs be used for each grade and subject area. This type of study will provide data that can be used to evaluate the content representativeness of items, sets of items that comprise an adaptive test for a student, and sets of items that comprise assessment targets, claims, or other levels of the test specifications. A contractor may propose a more general alignment study involving tasks that differ from those recommended here, which may be appropriate. However, the contractor should be required to demonstrate how the data will confirm the congruence between the sets of items that comprise an assessment for a student and the test specifications, as well as the degree to which the test items adequately represent the targeted cognitive skills. Although the adaptive nature of the summative assessments makes aggregating content validity results to a test "form" impossible, the representativeness of the most common sets of items taken by examinees, or a representative sample, could easily be studied (e.g., Crotts, Sireci, & Zenisky, 2012; Kaira & Sireci, 2010).

The content validity studies should also break out the results by item format. The summative assessments will include traditional selected-response items, technology-enhanced items, and performance tasks. Ideally, all item formats should have high ratings.

There is one drawback to the content validation/alignment methods discussed so far. By informing the SMEs of the CCSS measured by the items or of the assessment targets measured, they may exhibit a "confirmationist bias" or social desirability. That is, the SMEs may unconsciously rate the items more favorably than they actually perceive them, to please the researchers. One way around this problem is to have SMEs rate the *similarity* among pairs of test items and use multidimensional

scaling to analyze their data (D'Agostino, Karpinski, & Welsh, 2011; O'Neil et al., 2004; Sireci & Geisinger, 1992, 1995). However, this approach is not very common because it takes more time for SMEs to complete and involves more complex data analysis. A description of this method appears in Appendix C, should concerns about confirmationist bias/social desirability in evaluating test content arise.

Evaluating evidence-centered design. The evidence-centered design (ECD) underlying the development of the summative assessments specifies four claims and accompanying rationales in each subject area. These claims represent the cognitive models for each subject area. The assessment targets provide the evidence to support the claims, and the score reports represent the interpretation of the evidence. The content validity studies previously described could be extended to evaluate these three components of ECD in each subject area. The survey of curriculum specialists described earlier could include questions regarding the soundness of the claims and accompanying rationales in each subject area. Second, the studies involving ratings of items could be aggregated at the assessment target level to ensure that each target is represented by a sufficient number of items that are rated as measuring their intended CCSS well.

The third aspect of ECD, interpretation, should be evaluated through studies regarding the utility and comprehensibility of the summative assessment score reports. Ideas for these studies are described later in this report, in sections regarding validity evidence based on testing consequences. The idea here is to discover whether users of test reports interpret them correctly (Haertel, 1999), as well as if there are means for improving these score reports. It is assumed that studies of this kind will be done via piloting of the score reports. However, studies of the utility of the score reports should include ascertaining whether the information in the score reports is readily interpretable with respect to the intended claims.

Validity Studies Based on Internal Structure. Validity studies based on internal structure should be conducted to support the interpretations made on the basis of scores from the summative assessments. The scores reported should demonstrate adequate reliability and confirm the hypothesized “dimensionality” of the assessment. Studies in this area will involve analyzing the data from students’ responses to the items.

Dimensionality assessment. With respect to dimensionality, it is presumed that items comprising the summative assessments will be calibrated using unidimensional IRT models, which are the most common models in contemporary educational assessment. One straightforward way to assess the dimensionality of tests calibrated using IRT is *residual analysis* (Hambleton, 1989; Hambleton & Rovenelli, 1986). Residual analysis compares the probability of success on an item (predicted by the IRT model) for students of different proficiency levels to the actual success of students of different proficiency levels.

Two examples of residual analysis plots are presented in Figures 5 and 6. The small circles in each figure are “conditional p -values” and represent the proportion of students, within a certain test score interval, who correctly answered the item. That is, they are proportion-correct statistics, conditional on test score (actually, conditioned on the IRT estimate of true score, called θ). The vertical lines spreading from these conditional p -values illustrate the confidence intervals for the probability estimates based on the IRT model. The item displayed in Figure 5 displays good fit, in that the IRT model for this item essentially runs through the conditional p -values. The item displayed in Figure 6 does not fit well, as several of the conditional p -values are far off the item characteristic curve specified by the IRT model.

Inspection of residual plots is descriptive in nature, and there are statistical indices that can be used to flag items that do not fit the IRT model. Such analyses are important for the summative assessments, to make sure that the various item types used are all adequately fit by the IRT model. More importantly, however, summary statistics across all items can be used to evaluate the degree to which the IRT model fits the data for all items comprising an assessment, and hence the degree to

which the IRT assumption of unidimensionality holds (note that a lack of fit may indicate a problem other than multidimensionality). All of the aforementioned analyses can be conducted using customized software, or the free ResidPlots2 residual analysis software developed by Liang, Han, and Hambleton (2008, 2009).² The ResidPlots2 software allows users to simulate data that fit the IRT model, to gauge the degree to which the observed test data deviate from chance expectations, assuming the IRT model is true. This analysis can be useful for evaluating overall IRT model fit to the data. Further description of ResidPlots 2 appears in Appendix D.

It should be noted that most IRT software programs produce residual plots and statistical measures of fit, such as the chi-square statistic. If the Smarter Balanced assessments were calibrated using the Rasch model, the Infit and Outfit measures of item fit could also be used to evaluate IRT model fit (e.g., Linacre, 2004).³

² Available for free from the University of Massachusetts at <http://www.umass.edu/remf/software/residplots/>.

³ Both Infit and Outfit summarize the residuals between a student's observed pattern of responses to a set of items and the pattern predicted from the IRT model. The difference between the two measures is that the Infit measure weights items "closer" to a student's proficiency (theta) score more heavily than items further from the student's proficiency, whereas the Outfit statistic does not involve weighting. Each statistic represents a mean square error of the residuals and each has a standardized version.

Figure 5. IRT Residual Analysis Plot from ResidPlots-2 (good model fit)

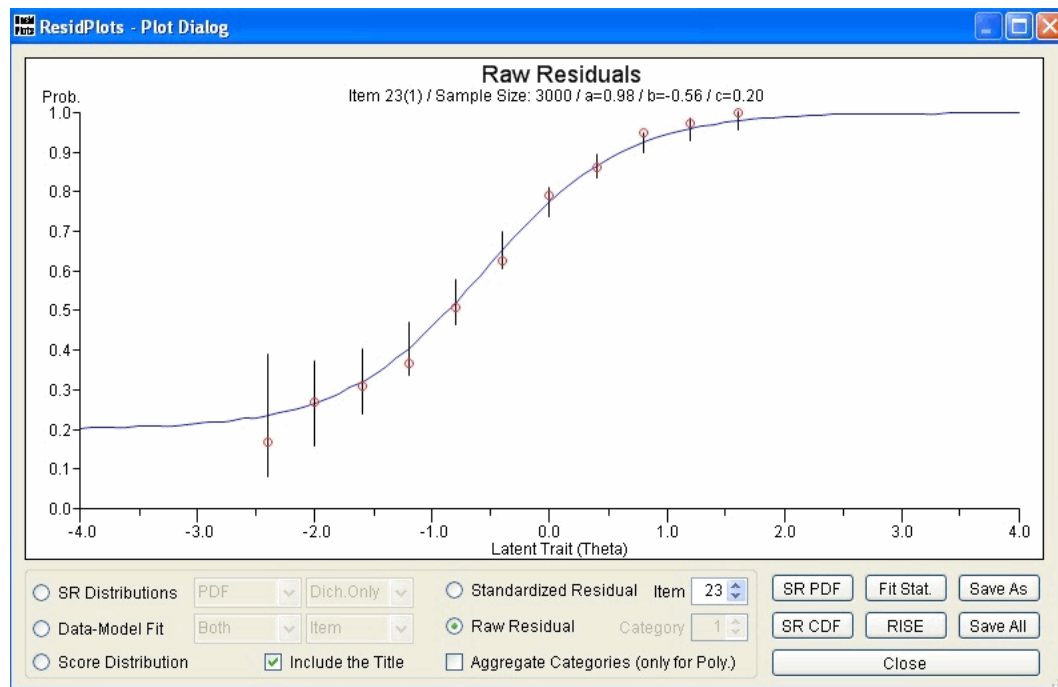
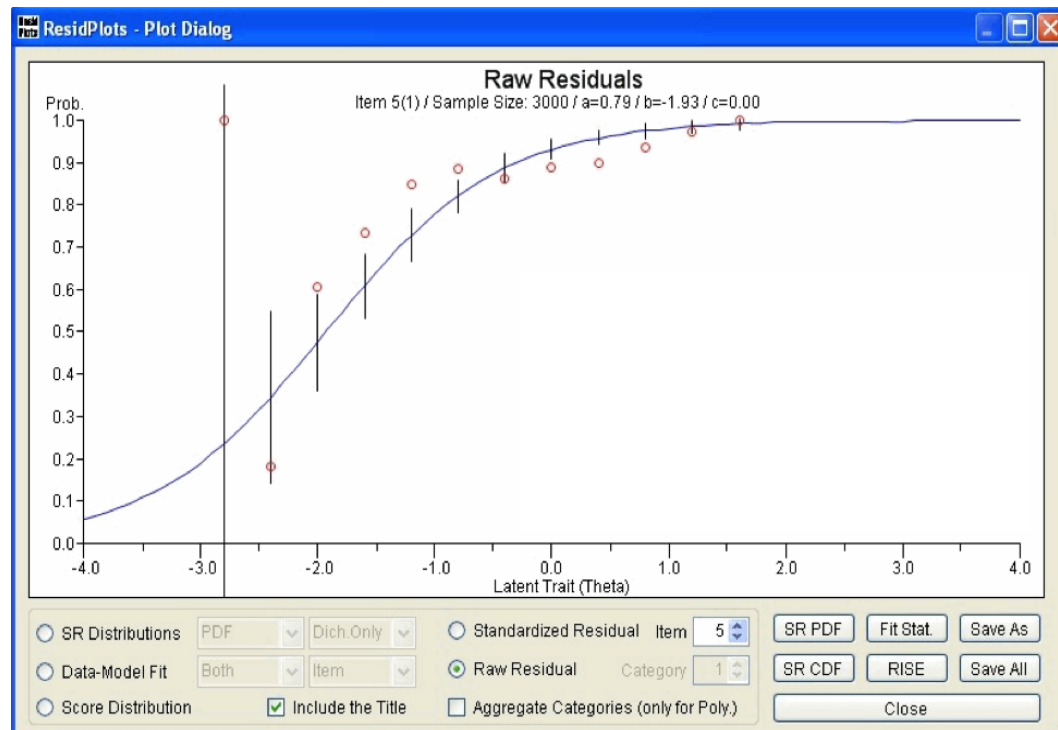


Figure 6. IRT Residual Analysis Plot from ResidPlots-2 (poor model fit)



There are more comprehensive methods for assessing the dimensionality of an educational assessment, such as exploratory and confirmatory factor analysis and multidimensional scaling (see Hattie, 1985, or Sireci, 1997, for reviews of methods). Some of these methods are recommended for validity studies related to other Smarter Balanced purposes. For purpose 1, which is focused on whether the assessments are valid and reliable measures of the CCSS, evaluating dimensionality via residual analysis should be sufficient. An advantage of IRT residual analysis is that it can be easily conducted on “incomplete” data sets that result from adaptive testing—that is, the student-by-item data file is incomplete in that not all students respond to all items. Such nonrandom, missing data is difficult to analyze using standard factor analytic procedures (cf. Sireci, Rogers, Swaminathan, Meara, & Robin, 2000).

Measurement precision. Purpose 1 for the summative assessments specifies reliable measures, which involve an analysis of the precision of the assessments. *Measurement precision* refers to the amount of error, or variation, expected in a student’s test score if the student were repeatedly tested. It is closely related to test score *reliability*, which is an estimate of the consistency or stability of the score. As described by Anastasi (1988):

Reliability refers to the consistency of scores obtained by the same persons when reexamined with the same test on different occasions or with different sets of equivalent items, or under other variable examining conditions. This concept of reliability underlies the computation of the *error of measurement* of a single score, whereby we can predict the range of fluctuation likely to occur in a single individual’s score as a result of irrelevant, chance factors. (p. 109)

Measurement precision is a broader term than *reliability* and refers to both estimates of score reliability and other descriptions of measurement error. A great deal of statistical theory has been developed to provide indices of the reliability of test scores as well as measures of measurement error throughout the test score scale. Classical test theory defines reliability as the squared correlation between observed test scores and their unbiased values (“true scores”). Reliability indices typically range from 0 to 1, with values of .80 or higher signifying test scores that are likely to be consistent from one test administration to the next.

Reliability indices are based on “classical” theories of testing. These estimates are reconceptualized in IRT, which characterizes measurement precision in terms of test information and conditional standard error. Therefore, the recommended measurement precision studies to support purpose 1 include estimates of score reliability (both coefficient alpha and stratified alpha, where relevant) and analysis of conditional standard errors of measurement based on IRT (e.g., test information functions and standard-error functions). Estimates of decision consistency, decision accuracy, and generalizability studies will be discussed in the sections related to other study purposes.

Validity Studies Based on Response Processes. The CCSS specify a wide range of knowledge and skills in each subject area. For example, two standards in high school geometry are:

Know precise definitions of angle, circle, perpendicular line, parallel line, and line segment, based on the undefined notions of point, line, distance along a line, and distance around a circular arc.

and

Construct an equilateral triangle, a square, and a regular hexagon inscribed in a circle. (NGA Center & CCSSO, 2010, p. 76)

The first standard represents a lower cognitive level of knowledge, while the second represents a higher level involving synthesis of several geometrical concepts. Evidence based on students' response processes could help validate that the summative assessment items are measuring the lower- and higher-order cognitive skills specified in the CCSS. One relatively easy study that could be done is an analysis of the amount of time it takes students to respond to items of various (purported) cognitive complexity. Students' response-time data should be readily available after the pilot tests, and the hypothesis that the items measuring higher-order skills will take more time for students to complete could be tested using analysis of variance (ANOVA).⁴ In addition, cognitive interviews or think-aloud studies could be conducted to best understand students' thought processes as they respond to items of varying cognitive complexity (Hamilton, 1994; Leighton, 2004).

Summative Assessment Purposes 2 and 3:

Provide valid, reliable, and fair information about whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA and mathematics to be on track for achieving college readiness.

and

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.

These two purpose statements reflect the fact 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 readiness. At grade 11, the achievement levels will include a “college and career readiness” category. Such classification decisions require validation. Validity evidence for these purposes should come from four sources—test content, internal structure, relations with external variables, and testing consequences. In addition, because these classification decisions represent achievement level standards, Kane's (1994) sources of validity evidence for standard setting—procedural, internal, and external—are also relevant. However, we note that Kane's external evidence overlaps considerably with validity evidence based on relations with external variables.

Summative assessment purposes 2 and 3 differ with respect to grade level, with the assessments prior to grade 11 being used to predict whether students are “on track” for college and career readiness, and the grade 11 assessments used for certifying certain academic aspects of college and career readiness. This difference involves somewhat different types of validation evidence. In particular, because there has been a great deal of work on assessing college readiness, there are more potential validation criteria for the grade 11 college readiness classification.

⁴ Note that response-time data are typically highly positively skewed, and so a natural log or similar transformation would be needed for this analysis.

Validating “On Track” Based on Content Validity Evidence. Being on track for college readiness implies acquisition of knowledge, and mastery of specific skills, thought to be important as students progress through elementary, middle, and high school. These specific knowledge and skills are stipulated in the CCSS. Therefore, the validity studies described earlier for purpose 1 are all relevant here. Essentially, the validity studies based on test content that were described for purpose 1 need to confirm that the summative assessments are targeting the correct CCSS and adequately represent these standards. However, such studies will not confirm that the CCSS actually contain the appropriate knowledge and skills to support college and career readiness. Rather, the CCSS would need to be reviewed to confirm that they contain the appropriate knowledge and skills that students need in order to be on track for college and career readiness.

One way to evaluate the appropriateness of the CCSS for determining whether students are on track for college and careers is to conduct a survey of state educators. At the postsecondary level, Conley, Drummond, Gonzalez, Rooseboom, and Stout (2011) conducted a national survey of postsecondary institutions to evaluate the degree to which the grade 11 and grade 12 CCSS contain the knowledge and skills associated with college readiness. They found that most (of almost 2,000) college professors rated these CCSS as highly important for readiness in their courses. A similar type of survey of educators in participating states would be helpful for evaluating the CCSS in ELA and math in grades 3 through 8. A major question motivating the survey would be: Are the CCSS in these grades appropriate for preparing students for college and careers?

In addition to these studies, it should be noted that studies involving validity evidence based on relations with other variables will also require validity evidence based on test content. For example, when Smarter Balanced assessment scores are compared with other test scores, the similarity of content across the two tests will need to be assessed.

Validating “On Track” Based on Internal Structure Evidence.

Decision consistency and decision accuracy studies. Given that purpose 2 involves the achievement level classification of “on track,” in addition to the measurement precision studies described earlier for purpose 1 (IRT residual analysis, reliability estimates, information functions, etc.), evidence that the *classifications* assigned to students are reliable is needed. Therefore, estimates of decision consistency (DC) and decision accuracy (DA) are needed, as are estimates of the precision of measurement around the “on track” cut score (i.e., conditional error of measurement at that point).

In essence, DC refers to the consistency of student classifications resulting either from two administrations of the same examination or from parallel forms of an examination. Thus, the concept is similar to reliability, but instead of consistency of a score, it refers to consistency of classifications across repeated testing. DA can be thought of as the extent to which the observed classifications of students agree with the students’ “true” classifications. Estimates of DA compare the classifications into which students are placed based on their test score with estimates of their true classifications. However, because students’ true proficiencies are never known, simulation studies or some type of split-half estimate are typically used to estimate DA.

There are several statistical approaches for estimating DA and DC. Livingston and Lewis (1995) introduced a method for estimating DC and DA based on a single administration of a test, using classical test theory. More recently, IRT-based methods have been proposed (Lee, 2008; Rudner, 2001, 2004) and are more common for IRT-based tests. Free software for estimating DC and DA for IRT-based tests, such as the Smarter Balanced summative assessments, is available (Lee, 2008),⁵ although some adjustments may need to be made for the adaptive test design. Another option would be the approach used by Hambleton and Han (2004), who estimated DA and DC by simulating data

⁵ This software, IRT-Class, is available for free from the University of Iowa via http://www.education.uiowa.edu/centers/docs/casma-software/IRT-CLASS_v2_0_for_PC.zip?sfvrsn=0.

based on IRT item parameter estimates, and by comparing the consistency of classification over simulated examinees.

Estimating the cut-score standard error. As Kane (1994, 2001) discussed, analysis of the expected amount of variability in the cut score resulting from a standard setting study should be considered in validating an achievement level standard. As part of the documentation for setting the “on track” standard and other achievement level standards on the summative assessments, estimates of cut-score variability should be provided. These descriptive statistics estimate the amount of change expected in a cut score if the study were replicated using different panelists, items, or standard setting methods. Sireci et al. (2009) provided examples of several different methods for evaluating the cut scores established on a grade 12 National Assessment of Educational Progress (NAEP) mathematics assessment. These methods range from simply computing the standard error of the mean across panelists to replicating the standard setting study using an independent standard setting panel.

For the “on track” college readiness standards below grade 11, estimates of cut-score variability should be documented, but should also be communicated to Smarter Balanced leadership *before* the cut scores are finalized. The specific estimates to be used are somewhat dependent on the standard setting method. Most methods involve cut-score recommendations for each panelist, and so the standard error of the panelist mean can be computed. Where multiple rounds of standard setting are conducted in a study, the variability (e.g., standard deviation, standard error of the mean) across rounds can be calculated, with the expectation that variability will decrease across rounds.⁶ When the panelists’ median cut score is used, standard errors for the median can be computed based on bootstrapping (e.g., Sireci et al., 2009) and other procedures.

A better estimate of cut-score reliability is based on the variability across independent standard setting panels. Brennan (2002) showed that when there are only two independent observations, such as two means from two separate standard setting studies, the standard error of the mean is

$$\hat{\sigma} = \frac{|X_1 - X_2|}{2}$$

where X_1 and X_2 are the means across panelists in the two standard setting studies. For Smarter Balanced summative assessments that involve high-stakes standards, we recommend that independent standard setting studies be conducted so that the variability across recommended cut scores can be estimated.

Validating “On Track” Based on Relations with External Variables. It is likely that one of the achievement level standards set on the ELA and Math summative assessments will be used as the “on track” designation in each grade level. For example, the “Proficient” standard in each grade might be used. Validating this specific score interpretation based on the relations of scores with other variables requires other measures of students’ mastery of grade-level knowledge and skills. Examples of external variables that could be used are teachers’ ratings of students’ preparedness for the next grade and other standardized assessments. Welch and Dunbar (2011), for example, explored the use of the Iowa Tests of Basic Skills (ITBS) for determining college readiness from grades 5 through 11. To accomplish this task, they first explored the relationship between the ITBS and the ACT composite scores for students who had taken the ITBS across grades and who had taken the ACT. The correlations between ITBS scores and the ACT ranged from .82 to .87 from grades 5 through 11. Next, for grade 11, they found the ITBS score that maximized classification

⁶ Although computing statistics such as the standard error of the mean is common in standard setting studies, when panelists discuss their ratings, the independence-of-observations assumption is violated, and so this estimate of variability probably underestimates the true variability across independent panelists.

congruence with the ACT college readiness benchmark score (their study involved students who took both assessments). Using the corresponding ITBS percentile rank scores at the lower grade levels, they found about an 80% accuracy rate for predicting the ACT benchmark. However, they suggested putting error bands around the “on track” benchmark, and if a student’s score was within the error band, the student could be considered on track.

In addition to the Welch and Dunbar (2011) study, both ACT and the College Board are using assessments at lower grade levels to assess college readiness. ACT has readiness benchmarks on its EXPLORE and PLAN assessments for grades 8 and 10, and the College Board recently introduced the ReadStep exam for grade 8 and has long used the PSAT in Grade 10. The ACT benchmarks for EXPLORE and PLAN were set by retrospective analysis of students who took EXPLORE, PLAN, and the ACT.

Another study that could be conducted is to have teachers classify their students regarding whether each student is prepared for the knowledge and skills to be taught at the next grade level. Although subsequent-grade-level preparedness is different from college readiness, it is likely that these two variables would be strongly related. Thus, the classification consistency between teachers’ ratings and students’ “on track” classifications could provide useful validity evidence. For this type of study, teachers would have to be familiar with the curricula taught in the subsequent grade. We also recommend gathering data on teachers’ *confidence* in the rating that they make for each student. Such data would be an important validity check before computing classification consistency and could be used to delete the data for teachers who were not confident in making their preparedness ratings for some or all students.

Validating “On Track” Based on Testing Consequences. Providing “on track” and other achievement level classifications for students in grades 3–8 is likely to have consequences for students, teachers, and instruction. At the student level, one potential negative consequence is promoting low academic self-esteem for students who are classified as below “on track.” Such negative feelings could lead to “self-fulfilling prophecies” where students begin to believe that they are not smart or not capable of graduating high school. Student surveys and tracking dropout rates over time (Rabinowitz, Zimmerman, & Sherman, 2001) are two ways that this and other consequences could be measured. The “on track” designation could also have the intended positive consequence of early identification and remediation of students classified as below “on track.” Therefore, following up on the instructional decisions that are made for these students is another area of study that would provide important validity evidence. Validity evidence for this purpose based on testing consequences should also involve gathering data from teachers via interviews, focus groups, or surveys to assess their perceived utility of these classifications and how it has affected their instruction. The consistency of these impressions and effects on instruction across grades should be studied.

Validating “On Track” Based on Procedural Evidence. Procedural evidence for standard setting refers to documentation and justification of all of the decisions and actions associated with a standard setting study. These decisions and actions were previously summarized in Table 5 (pp.20–21), and include selection of the standard setting panelists, justification of the standard setting method, training of panelists and other tasks associated with successful implementation of the method, analyzing the data, and assessing panelists’ confidence in their ratings and the process. Justification of the standard setting method will be important for the Smarter Balanced assessments, as some methods, such as the widely used Bookmark method, have been shown to have serious deficiencies (Davis-Becker, Buckendahl, & Gerrow, 2011; Reckase, 2006a, 2006b). Procedural evidence must be comprehensively documented, and should include surveys of panelists and others involved in the process. Standard setting reports for NAEP, such as those by ACT (2005a, 2005b, 2005c) are excellent examples of comprehensive documentation of standard setting that provides procedural, internal, and external validity evidence.

Validating College and Career Readiness Benchmarks

The third purpose statement for the summative assessments specifies college and career readiness. For the purposes of this research agenda, we assume that the knowledge and skills associated with college and career readiness have substantial overlap, as suggested by recent research (e.g., American Diploma Partnership, 2004; ACT, 2006), and so we focus on validating the college readiness benchmark. However, this assumption is based on convenience rather than research, since others have argued that the benchmarks for college and career readiness will be very different (Camara, in press; Loomis, 2011). Nevertheless, the methods described here for validating college readiness would carry over to the validation of career readiness, should appropriate external criteria for career readiness be identified.

Validating “College and Career Ready” Based on Content Validity Evidence. Up to this point, we have twice discussed validity evidence based on test content—first for purpose 1, and second with respect to students being “on track” for college readiness (purpose 2). The same studies apply here for validating the “college and career ready” inference based on the grade 11 summative assessments. This readiness designation implies acquisition of knowledge, and mastery of specific skills, considered necessary for success in college and careers and stipulated in the CCSS. Therefore, the content validity studies described earlier for purpose 1 are relevant here, and their findings should inform the validity argument for validating the college and career readiness standard. The additional evidence required for readiness is evidence that these standards are, in fact, the appropriate prerequisite skills in math and ELA that are needed to bypass remedial college courses and be ready to successfully begin postsecondary education or a career. The recent report by Conley et al. (2011) represents important evidence to support that assumption. Similarly, Vasavada, Carman, Hart, & Luisser (2010) found strong alignment between College Board assessments of college readiness and the CCSS.

Other validity evidence that is based on test content and that will be used in the validity argument for the college and career readiness determination includes content overlap (alignment) studies that will be done to gauge the similarity of knowledge and skills measured across the summative assessments and external assessments that are used to evaluate the readiness standards. Postsecondary admissions tests (e.g., ACT, SAT) and college placement tests (e.g., ACCUPLACER, AP, Compass) will be used in concurrent and predictive validity studies, and so the overlap of skills measured must be documented to properly interpret the results. The National Assessment Governing Board (NAGB) recently began a program of research in this area to set college and career benchmarks on the grade 12 NAEP assessments. Its research agenda began with comprehensive alignment studies that evaluated the overlap of NAEP and external assessments (Loomis, 2011; NAGB, 2010).

Validating “College and Career Ready” Based on Internal Structure Evidence. The previous descriptions of validity evidence based on internal structure for the “on track” student classification (i.e., estimates of DC and DA, review of the conditional standard error of measurement around the cut score, estimates of the standard error of the cut scores derived from the standard setting studies) are equally important for validating the college and career readiness classifications of students. These estimates and studies were described in previous sections, and so their descriptions are not repeated here.

Validating “College and Career Ready” Based on Relations with Other Variables. In considering validating the college readiness achievement level standards on the Smarter Balanced summative assessments, we focus on validity evidence based on relations to external variables because, as Camara (in press) pointed out, “Given the intended purposes of [college and career readiness] assessments, if performance levels and benchmarks are inconsistent with empirical data of performance in college and career-training programs, they will not only lack credibility but would raise concerns about the validity of the interpretive argument.”

A college- and career-ready standard implies that students who meet this standard have the prerequisite academic knowledge and skills to succeed in college or in a career. Given that there are currently existing standards for college readiness,⁷ the readiness classifications based on the Smarter Balanced summative assessments should be congruent with these other standards, assuming that these external standards accurately measure college readiness. The degree to which current college readiness benchmarks are consistent with the Smarter Balanced readiness standards needs to be studied. These studies could be used (a) to empirically set the Smarter Balanced readiness standards, (b) as part of the standard setting process, or (c) to validate the standards after they have been set by other means.

Validity evidence based on relations to other variables for the purpose of classifying students as college ready should involve both correlation/regression studies and classification consistency analyses. In these analyses, scores from the summative assessments will be correlated with, used as predictors of, and cross-tabulated with other measures of college readiness. To conduct these analyses, appropriate external measures must be identified, defined, and evaluated for validation purposes. In addition, different research designs should be considered. Design options include:

- Concurrent studies where students take both the summative assessments and external assessments;
- Predictive studies where students take the summative assessments and their future college performance is compared in retrospective fashion; and
- Embedded item designs where summative assessment items are embedded in other assessments of college success, and vice versa.

Defining “college success” is not straightforward, and so we recommend that several different variables be used, and studied, as outcome variables for college readiness. Camara (in press) listed seven criteria that have been or could be used for setting or evaluating college readiness benchmarks on Smarter Balanced or Partnership for Assessment of Readiness for College and Careers (PARCC) assessments. These are:

- Persistence to second year;
- Graduation or completion of a degree or certification program;
- Time to degree completion (e.g., 6 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.

Camara also noted that the most common criterion is college grades, either first-year grade point average (GPA) or grades in specific first-year courses. For example, in setting the college readiness benchmark on the ACT, grades in specific first-year courses were used (Allen & Sconing, 2005), but to set the same benchmark on the SAT, Wyatt, Kobrin, Wiley, Camara, and Proestler (2011) used first-year GPA.

⁷ We use *readiness* here to refer to the academic skills in math and reading, not the more general readiness criteria that include non-cognitive variables such as contextual skills and academic behaviors (Conley, 2007).

Current college readiness benchmarks set on educational tests. Several studies have been used to evaluate or set college readiness benchmarks on tests. Examples of testing programs that have set or evaluated college readiness benchmarks include:

- ACCUPLACER
- ACT
- Advanced Placement exams
- COMPASS
- Current statewide high school tests (end-of-course or graduation tests)
- Early Assessment Program (California)
- EXPLORE
- International assessments (e.g., PISA, TIMSS)
- International Baccalaureate
- NAEP
- PLAN
- PSAT/NMSQT
- ReadiStep

A recent report by NAGB (Fields & Parsad, 2012) found that the most common assessments used by postsecondary institutions to evaluate entering students for remedial courses in math were the ACT, SAT, ACCUPLACER (Elementary Algebra and College Level Math), and COMPASS (Algebra, College Algebra). For reading, the most common assessments were the ACT, SAT, ACCUPLACER (Reading Comprehension), ASSET (Reading Skills), and COMPASS (Reading).

Examples of some of the studies that have been done using these tests, the readiness standards that were set on each, and relevant citations are presented in Table 6. Camara (2012) described research in this area as consisting of three steps: First, determine the appropriate outcome variable for college success (e.g., first-year GPA). Second, determine the appropriate criterion of “success” on the outcome variable (e.g., 65% chance of a B-). Third, determine the appropriate probability of success. These steps will be important considerations in designing validity studies for the Smarter Balanced summative assessments.

Table 6. Current College Readiness Benchmarks

Test	Criterion	Benchmark	Comments/Citations
ACT English	.75 probability of C and .50 probability of B	18	Allen & Sconing (2005)
ACT Reading		21	
ACT Math		22	
SAT Composite	.65 probability of B- in first-year GPA	1550	Wyatt et al. (2011)
SAT-Quantitative		500	
SAT-Reading		500	
SAT-Writing		500	
Advanced Placement (AP)		Score of 3	Relevant tests include Calculus AB, Calculus BC, English Language & Composition, English Literature & Composition, and Statistics.
COMPASS	.75 probability of C and .50 probability of B	77 (English), 52 (Math)	ACT (2010)
EXPLORE	.75 probability of C and .50 probability of B	13 (English), 17 (Math)	ACT (2010)
PLAN		15 (English), 19 (Math)	ACT (2010)

The studies reported in Table 6 primarily used regression methods to find the test score that best distinguished students who met or did not meet some operationally defined criterion of college success.⁸ For the ACT research, the criterion used was the test score associated with a .75 probability of earning a C or a .50 probability of earning a B in specific college courses (e.g., English composition, college algebra). For the SAT research, the criterion used was the test score associated with a .65 probability of earning an overall first-year GPA of B- (2.67). The ACT studies used linear regression, whereas the SAT studies used logistic regression. The SAT studies also included validity evidence based on external variables, specifically rigor of high school courses, AP exam scores, and high school GPA, to support the SAT readiness benchmarks (Wyatt et al., 2011). In addition to the studies reported in Table 6, Fields and Parsad (2012) conducted a comprehensive survey of cutoff scores on postsecondary math and reading placement tests. The mean cutoff scores, and the variability in these scores across institutions, were reported. These mean cutoff scores could be used as validation criteria for the Smarter Balanced college readiness standards. Other readiness criteria include specific cutoff scores used by state university systems (e.g., California and Texas have readiness criteria based on the ACT, the SAT, and in-state assessments), and the International Baccalaureate exams (compensatory score of 24 across six assessments).

⁸ Equipercetile equating could also be used, and may be preferable in some situations.

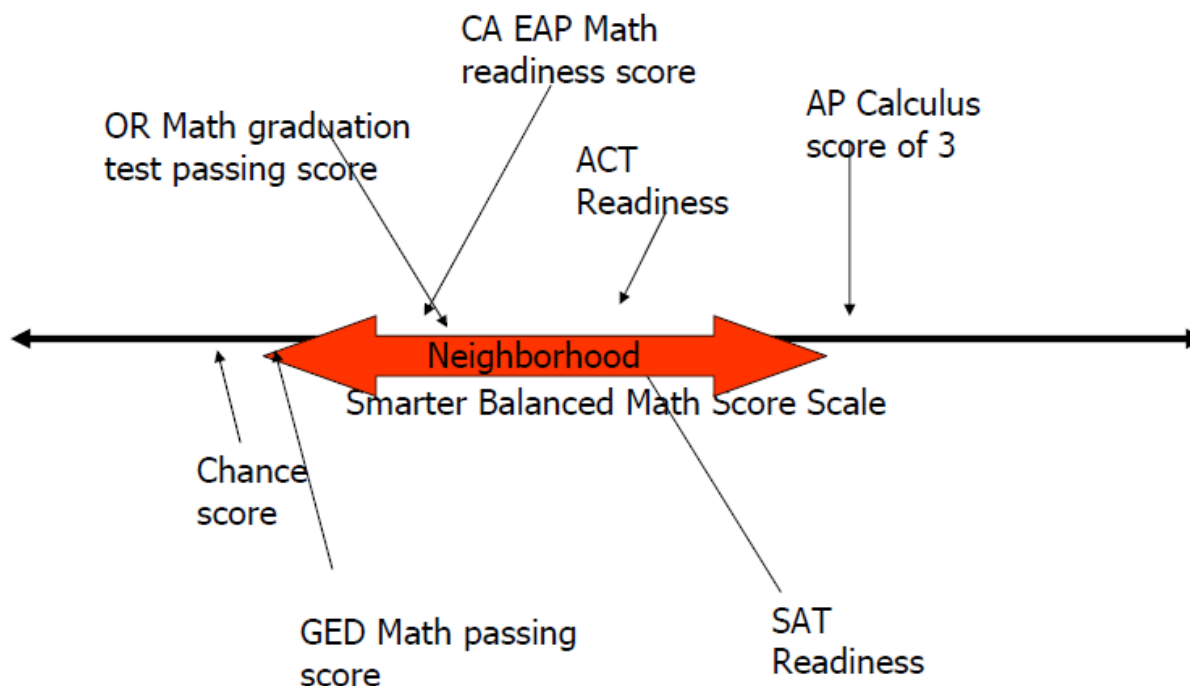
In addition to establishing college readiness benchmarks on admissions tests, research has also been conducted to see how these readiness benchmarks could inform setting readiness standards on other assessments. For example, the Texas Education Agency commissioned a series of studies to set and evaluate college readiness standards using the State of Texas Assessments of Academic Readiness (STAAR). In fact, in establishing the new STAAR tests, the Texas legislature legislated that “validity studies be conducted to evaluate the empirical links between student performance on the STAAR assessments and specific assessments measuring similar constructs, and that these links be used to inform the standard-setting process” (LaSalle et al., 2012, p. 2). These studies are particularly relevant to Smarter Balanced because the STAAR assessments involve on-target readiness standards below high school and certifying college readiness at the high school level.

Rather than directly using external assessments to set readiness benchmarks on the STAAR exams, Texas used external data to set “landmarks,” or cut points, on the STAAR score scale that corresponded to important cut scores on the external assessments. Examples of external assessments that were used for this purpose included the previous statewide exams in Texas, a placement test used at the University of Texas, the ACT and SAT benchmarks, and the ACCUPLACER Elementary Algebra exam. For the previous statewide end-of-course tests, equipercentile linking was used to establish concordance tables across pairs of tests. For the readiness benchmarks established on the external assessments, logistic or linear regression was used to “map” the external benchmarks onto the STAAR score scales. Linear regression was also used to set other landmarks based on high school course grades (e.g., B or better) and probability of success in a relevant college course (e.g., C or better in college algebra). See Keng, Murphy, and Gaertner (2012) for a more complete description of these studies.

Based on several studies of these external criteria, “landmarks,” or benchmarks, were established on the STAAR score scale, and these landmarks were used to establish “neighborhoods” within which it seemed reasonable (to the policymakers who reviewed these results) to set the college readiness standard and other standards. The score scale annotated with the landmarks and neighborhoods was used to encourage standard setting panelists to set their standards within the neighborhoods, since the score scale range defined by each neighborhood contained the external readiness standards and other relevant information that would support the standard set in that range. Keng et al. (2012) described this process as “evidence-based standard setting” (p. 4; see also O’Malley, Keng, & Miles, 2012).

A fictitious example of how external data could be used to inform the college and career readiness standard setting process using neighborhoods based on external data is presented in Figure 7. In this figure, test scores related to college readiness from two states (California and Oregon), the ACT and SAT readiness benchmarks, and the passing score for the GED Math test are all mapped onto the score scale for the grade 11 Smarter Balanced summative math assessment. The score corresponding to chance performance is also indicated. Using external data in this way can build validation criteria into the standard setting process.

Figure 7. Example of Using External Data to Establish a Reasonable Interval (Neighborhood) for Standard Setting



Recommended studies based on relations to external variables. The previous section described some options for conducting validity studies based on relations to external variables and summarized some of the research that has already been done in this area. To relate current college readiness standards and other pertinent information to the grade 11 Smarter Balanced summative assessments, three types of studies are possible. The first two types of studies are concurrent validity studies. In the first variation, students would take both Smarter Balanced and external assessments at around the same point in time. For example, grade 11 students could take the Smarter Balanced summative assessments, or a subset of items from them (e.g., in the pilot study), and the SAT or ACT, at a reasonable point in time (e.g., March). Regression or equipercentile methods could be used to determine the Smarter Balanced scores that corresponded to the SAT or ACT readiness benchmarks. The second type of concurrent validity study would involve college students taking Smarter Balanced assessments (or subsets of items) near the end of a relevant course, and their final course grades could be used as the validation criterion. The Smarter Balanced scores that are associated with the pre-established readiness criterion (e.g., grade of B-) could be established via regression or equipercentile procedures, or probability tables could be set up to relate the Smarter Balanced scores to specific grades. The third type of study that could be conducted would be a retrospective study where students who took the Smarter Balanced assessments would be followed longitudinally to see how they perform in college (see, for example, D'Agostino & Bonner, 2009).

Threats to the validity of these studies include differential motivation effects across the Smarter Balanced and external assessments, potentially non-representative samples of students due to the self-selection of external assessments, and a lack of overlap in the constructs measured by the Smarter Balanced and external assessments. Different grading standards and different admissions standards across colleges and universities, and across different types of institutions (public, private,

two-year, four-year) also present problems. Nevertheless, these issues can be considered and discussed when interpreting the results. Surveys or interviews of students participating in these studies could help understand these students' motivation to do well (Haertel, 1999).

The most practical course of action to gather external data to validate the Smarter Balanced college readiness standards is to take advantage of tests already taken by grade 11 students, such as the ACT, SAT, and AP exams, and relate them to their scores on the summative assessments.

Supplementary studies would need to evaluate the content overlap of these assessments and students' motivation to do well on the Smarter Balanced assessments. Assuming sufficient content overlap and motivation, benchmarks can be set to inform the establishment of the college readiness standards on the Smarter Balanced assessments (as done in Keng et al., 2012), and longitudinal analysis can be done at a later point in time to evaluate the standards and possibly revise them if necessary. The key information to gather is the degree to which students who reached the Smarter Balanced readiness standards were successful in college. Camara and Quenemoen (2012) suggested that the decision consistency of the ready/not-ready and successful/not successful in college classifications should be broken down across different types of institutions.

It is likely that data-sharing agreements that maintain student anonymity can be worked out between the Consortium and external examination programs, such as ACT and the College Board, and among state colleges and universities within the Consortium. In addition, as Camara and Quenemoen (2012) point out, the National Student Clearinghouse maintains enrollment records for a vast majority of postsecondary institutions and can be used to track retention and graduation rates that will be useful for evaluating the readiness standards. The percentages of students who are "Proficient" on the grade 12 NAEP Math and Reading assessments will also be evaluated with respect to the percentages of students who are classified as "college ready" on the respective Smarter Balanced assessments. Should the NAEP grade 12 results ever be reported at the state level, within-state NAEP/Smarter Balanced comparisons would be informative.

Validating "College and Career Ready" Based on Testing Consequences. The college and career readiness standard on the Smarter Balanced summative assessments is intentionally integrated with the "on track" standards set at the lower grade levels. The intended consequence of this system is better preparation of students so that they are prepared for college or careers by the time they graduate high school. This intended consequence can be measured by analyzing trends in college completion and remedial course enrollments over time, and by surveying secondary and postsecondary educators about students' proficiencies. However, validity evidence for the college and career readiness designation should also investigate unintended consequences, such as unanticipated changes in instruction, diminished morale among teachers and students, and increased pressure on students that may lead to dropout, or to pursuing college majors and careers that are less challenging. To evaluate these potential consequences, teacher surveys of enacted curriculum, student surveys of career aspirations, and psychological assessments of anxiety and academic self-concept could be conducted.

The recommended studies based on testing consequences that will target the college and career readiness purposes should include teacher surveys regarding changes in student achievement and preparedness over time and changes in teachers' instruction over time. We also recommend that students be surveyed regarding college and career aspirations. Student and teacher samples that are representative at the state level would suffice for these studies. If time and resources permit, assessing the anxiety levels of students regarding their likelihood of obtaining college or career readiness, and their academic self-concept, would also be helpful. Validity evidence based on the consequences of the college and career readiness standard should also involve analysis of secondary and postsecondary enrollment and persistence, changes in course-taking patterns over time, and teacher retention for teachers in math and ELA.

Summative Assessment Purpose 4:

Provide valid, reliable, and fair information about students' annual progress toward college and career readiness in ELA and mathematics.

As indicated in Table 1, validity evidence to support the use of the summative assessments for providing information about students' annual progress should be based on test content, internal structure, relations with external variables, and testing consequences. Studies related to test content need to evaluate the degree to which similar standards are measured across grades and the consistency of the construct across grades. Studies based on internal structure should evaluate the validity of the vertical scale used to measure progress over time. Studies involving relations with external variables are needed to confirm that the progress observed on the Smarter Balanced scale is mirrored by other measures of academic achievement. Finally, studies based on testing consequences should confirm that the measures of annual progress have a positive effect on instruction and student learning.

The most straightforward way to measure changes in students' proficiencies over time is to have scores from assessments at different points in time on a common scale. The physical analogy is the bathroom scale that remains unchanged across different measurements of weights. Sometimes, however, even the bathroom scale needs to be recalibrated to confirm the zero point. With educational assessments, it is difficult to put scores from assessments at different time periods on the same scale, because the items administered to students at different points in time are not the same. At this juncture, the Smarter Balanced summative assessments are planned to be vertically equated across grades, which means that a single score scale will span the grades. A vertical scale facilitates measuring changes in students' performance over time (Briggs, 2012; Kolen, 2011; Patz, 2007). However, it is difficult to create a valid vertical scale. Challenges to vertical scaling include changes in the construct of math or ELA across grades, and differences in when material is taught across grades and schools (Tong & Kolen, 2007). Therefore, validity evidence to support measuring students' progress toward college and career readiness should involve evaluation of the vertical scale across grades.

Validity Studies Based on Test Content. Evaluations of the content measured across grades will be an important source of evidence for validating the appropriateness of the vertical scale for measuring students' progress. First, this evaluation should assess whether there is overlap among the CCSS measured across adjacent grades (Patz, 2007). Next, the evaluation should review the common items that are used to form the vertical links across grades. SMEs should be asked whether the linking items are relevant to students in both grades and if they adequately represent the expected learning progressions. The content review should also assess the degree to which a common construct can be considered to hold across grades, or at least across adjacent grades. For example, do the anchor items that are used across grades measure CCSS that are appropriate for each grade?

Validity Studies Based on Internal Structure. Most of the studies that should be conducted to evaluate the validity of the vertical scales underlying the summative assessments can be categorized as evidence of internal structure. These studies include dimensionality analyses and evaluation of item statistics, mean scores, and score distributions across grades.

Dimensionality analyses. One important area of study is evaluation of the dimensionality of the assessment data, and of the degree to which the dimensionality is consistent across grades, or at least across adjacent grades. For example, if a single dimension is hypothesized to exist across grades, the degree to which the data for each grade are unidimensional, and the degree to which the same dimension holds across grades, should be studied. One way to conduct this analysis is using IRT residual analysis, as suggested earlier. The added layer of analysis would be evaluating the consistency of the fit across grades. Kolen (2011) noted that "even if the unidimensionality assumption does not strictly hold, the IRT model might provide an adequate enough summary of the

data that the vertical scale is still useful” (p. 12). Other dimensionality assessment procedures, such as confirmatory factor analysis or bifactor analysis, could also be useful.

The incomplete student-by-item data matrix that results from adaptive testing can cause problems for many dimensionality assessment procedures, such as exploratory and confirmatory factor analysis. Thus, assessing the dimensionality within and across grades within an IRT framework is probably most practical. In addition to residual analysis, both unidimensional and multidimensional IRT models can be fit to the data, and the difference between models can be tested for significant and practical improvement in fit to the data (Bock, Gibbons, & Muraki, 1988; Sireci, 1997). For items that are dichotomously scored, this analysis can be conducted using the TESTFACT software (Wilson, Wood, & Gibbons, 1991). To assess multidimensionality using both dichotomous and polytomous items, some specialized software may be needed.

Analysis of statistics across grades. The establishment of a vertical scale implies an increase in the difficulty of the assessments as grade increases and higher proficiency of students in higher grades relative to lower grades. At the item level, it is assumed that students at a higher grade level will have a higher probability of correctly answering an item than students at a lower grade level. These assumptions can be checked to evaluate the validity of the scale. Factors such as when students are taught specific knowledge and skills (i.e., opportunity to learn) and difference in time between instruction and assessment can cause “reversals” where students at higher grade levels perform worse than students at lower grade levels. Such reversals can be a problem when a common item approach is used to link the assessments across grade levels. Therefore, an additional study is a comparison of where the items “land” on the vertical scale versus the grade levels for which they were written. For example, if items written for a grade 6 assessment have IRT difficulty estimates that put them in the general range of grade 5 or grade 7 items, there will be a disconnect between the intended content at each grade level and the actual scale properties.

Kolen (2011) and Patz (2007) suggested several analyses that could be used to evaluate the validity of a vertical scale (see also Kolen & Brennan, 2004). These analyses include:

- correlation of item difficulties across grade levels
- a progression in test difficulty of test characteristic curves across grades
- analysis of item difficulties across grades
- comparison of mean scores across grades
- comparison of scale scores associated with proficiency levels across grades
- comparison of overlap of proficiency distributions across grades
- comparison of variability in test scores within and across grades

Validity evidence for vertical scales that are appropriate for measuring students’ annual progress would include a lack of reversals of item difficulties across grades, anticipated separation of means and proficiency distributions across grades, and sensible patterns of variability within and across grades. With respect to comparison of score means across grades, Patz (2007) suggested, “For sufficiently large and diverse samples of students, scale score means would be expected to increase with grade level, and the pattern of increase would be expected to be somewhat regular and not erratic” (pp. 17–18).

With respect to evaluating patterns of variability, Kolen (2011) noted:

Within grade variability indices typically are either similar across grades or increase as grade increases. Either of these patterns seems reasonable. Sometimes within grade variability indices decrease substantially as grade increases, which is sometimes referred to as *scale shrinkage*. Scale shrinkage can be indicative of

problems with IRT parameter estimation, in which case the vertical scaling procedures might need to be adjusted or the scale abandoned. (p. 12)

In considering establishing a vertical scale for PARCC, Kolen noted, “PARCC might decide, based on the construct being assessed, that an acceptable vertical scale should display increasing mean scores from year to year, that the amount of growth is decelerating, and that the within grade variability is either approximately equal across grades or is increasing from grade to grade” (p. 21). These evaluation criteria are applicable to evaluation of the Smarter Balanced vertical scale for the summative assessments.

In addition to analyses of item statistics and test scores across grades, Briggs (in press) claims that vertical scales should be validated by demonstrating that they possess interval scale level properties. This idea is new and has not seen wide application, but Briggs suggests the use of additive conjoint measurement to determine whether vertical scales have equal-interval properties, which he considers necessary for valid measurement of students’ annual progress

In addition to the previously mentioned studies, analyses of *item parameter drift* over time should also be conducted. These analyses involve recalibrating IRT item parameters in subsequent years and comparing them to their estimates in prior years. Such analyses could improve the anchors used in equating across years by eliminating anomalous items, or could identify items that have been compromised (i.e., security problems).

Validity Studies Based on Relations to Other Variables. To confirm that the summative assessments provide valid information about students’ annual progress in math and ELA, it would be good to compare students’ progress on these assessments with other measures of their achievement over the same time period. At a macro level, the aggregated progress of students over time could be compared to changes of students within a state on the NAEP math and reading assessments. On an individual student level, progress on the Smarter Balanced assessments could be compared to other standardized assessments that are on a vertical scale, such as the ITBS or the Measures of Academic Proficiency (Northwest Evaluation Association, 2005).

In addition to concurrent validity evidence based on other tests, the degree to which the summative assessments are sensitive to instruction could also be studied to evaluate the degree to which the tests measure students’ annual progress. Teachers who more fully implement the CCSS into their instruction should have students who make greater progress on the summative assessments. D’Agostino, Welsh, and Corson (2007), for example, measured the degree to which teachers emphasized state academic standards in their teaching and compared these measures to students’ performance on the statewide test. They found a modest but positive relationship. A similar strategy could be implemented to evaluate the patterns of progress noted across classes on the summative assessments. Another way in which external data can inform the validation of the summative assessments as a progress measure is to have teachers rate the math and ELA progress made by their students within a year, and compare it to their progress as measured by the Smarter Balanced score scales.

Validity Studies Based on Testing Consequences. The summative assessments are supposed to provide information regarding students’ annual progress so that their progress toward college and career readiness can be ascertained. If adequate progress is not found, it is likely that instructional changes will be made to support improved progress. Thus, validity evidence based on testing consequences should include surveys or interviews of teachers to understand the degree to which they find estimates of students’ progress helpful for targeting instruction to individual students and to their classes in general. In addition, if progress measures are used to alter the instruction for a student—for example, placing the student in supplementary instruction or an after-school program—the degree to which these actions are associated with improved progress should be studied (Shepard, 1993).

Another important study of testing consequences related to measuring progress is the degree to which progress is similar across subgroups of students. If students from different ethnic backgrounds, socioeconomic statuses (SES), or disability statuses are progressing at different rates, the reasons for such differential progress should be studied. It may be that students who initially perform low on the assessments have more opportunity to exhibit progress. In any case, patterns of progress across subgroups should be studied to ascertain whether these patterns are expected given the student characteristics, or if they reflect some insensitivity of the assessments to properly capture progress or some type of deficiency in the scale properties.

Summative Assessment Purpose 5:

Provide valid, reliable, and fair information about how instruction can be improved at the classroom, school, district, and state levels.

As indicated in Table 1, for the Summative Assessments to provide information that will improve instruction, the content of the assessment must adequately measure the intended CCSS, and teachers, administrators, and other educators must appropriately act upon this information to tailor instruction accordingly. The validity studies based on test content that were described earlier for purposes 1 through 4, and the studies of testing consequences that were described for purposes 2 through 4, would all provide evidence regarding the degree to which the assessment results are instructionally relevant. The gathering of additional validity evidence to support purpose 5 will be similar to the studies suggested later in this report for the interim assessments and formative assessment resources, because these components are designed to work together to improve instruction. Many of these studies fall under the category of validity evidence based on testing consequences; one study based on relations to other variables, which was already mentioned with respect to purpose 4 (a study of sensitivity of the summative assessments to instruction), is also relevant to purpose 5.

As noted earlier, teachers who more fully implement the CCSS into their instruction should have students who make greater progress on the summative assessments (D'Agostino et al., 2007).

Validity Studies Based on Testing Consequences. The provision of summative assessment information to improve instruction will most likely come from the score reports associated with these assessments. Therefore, the evaluation of testing consequences relative to this purpose will focus largely on the utility of these score reports. An analysis of classroom artifacts will also provide important evidence, as will the types of surveys, interviews, and focus groups associated with the studies mentioned earlier for purposes 1 through 4.

Studies on effectiveness of summative assessment score reports. According to the score reporting RFP (RFP-15), Smarter Balanced has planned a wide and comprehensive variety of score reports to support purpose 5. There will be both static score reports and dynamic score reports that are interactive. Summative assessment results will be reported at the total score and claim levels for both subject areas, and reports will be available for both individual students and aggregate groups. The comprehensive nature of these reports, and their online access and variety, should provide actionable data to improve instruction at the classroom, school, district, and state levels. Research studies should be conducted to confirm that these intended consequences are occurring.

RFP-15 requires gathering feedback from potential users as score reports are being developed. Documentation regarding these reports should be reviewed to see what changes were made on the basis of this feedback. In addition, once the reports are operational, studies should be conducted to ascertain how well teachers, administrators, parents, students, and other stakeholders (e.g., legislators, journalists) understand the reports and find them useful. These studies should include surveys, focus groups, and interviews. In addition to gathering stakeholders' impressions of the reports, their understanding of the information contained in the reports should be tested (Wainer, Hambleton, & Meara, 1999). The actions that teachers take based on the score reports should also

be documented and evaluated for appropriateness (Bennett, 2010). In addition to assessing users' understanding and use of the reports, surveys should also be used to inquire about ease of navigating the system, timeliness of data, and additional features that users would like to see.

Analyses of usage statistics should also be conducted to determine the most popular reports and to confirm that all reports created are being used. The different types of reports that users create should also be reviewed. The most commonly used and least commonly used reports could be targeted for discussion in focus groups to (a) ensure that users are making appropriate inferences from the reports, (b) ensure that taking appropriate actions based on the reports, and (c) discover how the least-accessed reports could be improved to make them more useful, or to make users aware of them.

To maximize utility of the reports, users or "data coaches" should be trained on how to access them and use them. In fact, the Peer Review Guidance (U.S. Department of Education, 2009b) stated that "Training on interpretation of results is required [and] must provide evidence on how educators can interpret results and then use them for proper decision making" (p. 69). Thus, the effectiveness of the training should also be evaluated.

Studies of textbooks and classroom artifacts. Another way in which the effects of the summative assessments on instruction can be evaluated is by looking at changes in textbooks and instructional practices before, during, and after implementation of the assessments. In addition to the surveys and interviews previously discussed, classroom artifacts such as lesson plans, student handouts, classroom assessments, homework, syllabi, and teacher logs (e.g., Silk, Silver, Amerian, Nishimura, & Boscardin, 2009; Tomlinson & Fortenberry, 2008) should be studied.

Summative Assessment Purpose 6:

Provide valid, reliable, and fair information about students' ELA and mathematics proficiencies for federal accountability purposes and potentially for state and local accountability systems.

Results from the summative assessments will include scale scores at the total score and claim levels, and achievement level classifications in each subject area. The achievement level results could be used as they are currently employed in statewide testing programs for federal accountability purposes under NCLB. In addition, students' progress over time could be used in growth models for other accountability purposes, some of which may be for federal accountability and some at the state or local levels. The Smarter Balanced principle of "responsible flexibility" (Smarter Balanced, 2010, p. 5) is consistent with the idea of providing valid, reliable, and fair information that can be used for federal accountability in uniform fashion across all participating states, but also allows for states to use information from the summative assessments in their statewide and local accountability systems.

Smarter Balanced cannot assume the responsibility for validating all of the potential uses of the summative assessments at the state and local levels, but the responsibility for validating accountability at the federal level should be included in the research agenda. In particular, the metric of "percent proficient" at the total student population level and at the subgroup level should be validated, as well as any other aggregate statistics used for federal accountability.

Percent proficient is currently a primary accountability criterion in NCLB, which also requires states to set at least three proficiency levels. In considering the reporting of achievement level results in California, a technical advisory committee led by Lee Cronbach (Select Committee, 1994/1995) recommended that (a) the percent *above* cut points be reported, rather than percents at proficiency levels; (b) only one percent above cut points, or two at most, rather than percent above cut points for all proficiency levels, be reported; and (c) standard errors for percent above cut points be reported (Yen, 1997). The first two recommendations were suggested to reduce confusion in reporting scores

to the public. The third recommendation is standard practice in reporting scores for accountability or other purposes.

The provision of valid, reliable, and fair information has been covered in the previous purpose statements, through the various studies involving test content, internal structure, relations to other variables, response processes, and testing consequences. The additional studies needed to validate the accountability uses of Smarter Balanced summative assessment scores are studies involving the reliability and validity of *aggregate* scores used for accountability. Of particular importance is the reliability of aggregate scores.

Studies Evaluating the Reliability/Precision of Aggregate Scores. Individual schools will be one aggregate level of analysis in federal accountability, and so the reliability or error associated with school-level results will need to be estimated as part of the validity research agenda. If accountability results will be reported at more micro levels, such as classrooms, the reliability or error associated with those results would need to be estimated as well. The goals of the measurement precision studies to be done here are to provide an estimate of the error inherent in any aggregate scores that are reported for the summative assessments and to judge the utility of the information given the estimates of error. It is possible that these studies will support the use of the summative assessment data for accountability purposes at some levels (e.g., districts) but not others (e.g., schools), because of the increased sampling error associated with smaller numbers of students.

Several methods have been proposed to estimate the reliability, or standard errors, associated with aggregate scores from statewide assessments. Yen (1997) used generalizability theory (G-theory) to estimate the reliability of school-level results for percent-above-cut statistics associated with the Maryland State Performance Assessment program and evaluated a criterion of achieving a standard error, of these percents, of 2.5% or less. She concluded that was an unrealistic criterion for performance assessments in a single subject area, but could be reached when evaluating a composite across subject areas. Her study illustrated the utility of G-theory for estimating standard errors for aggregate statistics, regardless of the item formats that are used.

Hill and DePascale (2003) asserted that the reliability of decisions at the school level should be evaluated from a decision consistency perspective. That is, if the assessment were repeated, would a school receive the same (AYP) classification? Hill and DePascale (2002) listed four methods for estimating school classification consistency. The first, “direct computation,” is based on errors associated with each single classification and “uses areas under the normal curve to determine the probability of a correct classification” (p. 4). The second method is based on randomly dividing the students in a school into two groups and calculating the accountability statistics on each half. The third method involves randomly selecting (with replacement) multiple samples from a school, and the fourth method involves Monte Carlo simulation, where the parameters for a school are estimated and then random draws of students are made. In all four methods, the consistencies in schools’ classifications are evaluated. Hill and DePascale recommend using at least two methods to offset the disadvantage of any single method.

Regardless of the method used to estimate the reliability of or error associated with aggregate summative assessment statistics used for accountability, it is important that the estimates address both measurement error and sampling error (Hill & DePascale, 2003; Linn, Baker, & Betebenner, 2002), as do the aforementioned approaches by Yen (1997) and Hill and DePascale (2002, 2003).

Simulation or empirical studies should also be conducted to evaluate the impact of factors outside of a school’s control (or outside of the control of whatever the unit of inference is, such as a teacher) on the accountability results. For example, the inference made about a district or a school should not be statistically biased based on the number of students, the number of subgroups of students, or other factors beyond instruction. By estimating and using standard errors associated with aggregate scores when making accountability decisions, the validity of those decisions will be enhanced.

Simulation and other studies could also be used to inform accountability decisions such as how many years of data should be used to evaluate a district, school, or other unit of interest.

The degree to which derivative measures of summative assessment scores, such as “growth” measures, will be used in accountability systems is not known at the time of this writing. Any derivative measures would need to demonstrate evidence of reliability and validity. The Standards made this point when discussing what today might be considered a “growth” score: “When change or gain scores are used, the definition of such scores should be made explicit, and their technical qualities should be reported” (AERA et al., 1999, p. 167). Unfortunately, many of the current score derivatives, such as growth percentiles and value-added scores for teachers, have not been widely studied. As Brennan (2011) lamented, “to the best of my knowledge the subject of error variances and measures of precision for measures of growth is largely uncharted territory” (pp. 16–17).

Validity Studies Based on Relations to Other Variables. The use of summative assessment results for federal accountability purposes will certainly involve the use of achievement level results. In addition to the reliability studies previously mentioned, the previously mentioned studies supporting the use of achievement level standards are also relevant. However, additional studies are needed to support the utility of aggregate results based on achievement level results. For example, are the schools that are identified as not making adequate progress, based on percentages of “Proficient” or “on track” students, really the schools that should be flagged? Studies that could be designed to answer this question include using other measures of student achievement to classify schools into performance categories, and single-case studies where schools identified as over- or underperforming are carefully reviewed to evaluate the classification.

With respect to other measures of student achievement, at the high school level, changes in summative assessment scores for a school could be compared with the school’s changes in scores on AP and college admissions tests. Perhaps student fees for these admissions tests could be paid for to remove the self-selection problem. At the middle school level, ACT’s and the College Board’s assessments for younger students (EXPLORE, PLAN, Readiness) could be used.

Validity Studies Based on Testing Consequences. The use of test scores for accountability has been accused of causing many problems, such as decreased teacher morale, increased pressure on students, and narrowing of the curriculum. As described earlier for purposes 1 through 3, these criticisms could be studied using comprehensive surveys of students and teachers, both before and after the implementation of the summative assessments. Surveys could be used to understand the effects on students (e.g., anxiety, educational aspirations), teachers (morale, retention, movement into non-tested subject areas, instruction), administrators (e.g., teacher recruitment and retention, effectiveness of school improvement), and parents (e.g., observations of their child, school choice). Teacher retention rates and teachers’ movement into non-tested subject areas should also be tracked and studied.

Summative Assessment Purpose 7:

Provide valid, reliable, and fair information about students’ achievement in ELA and mathematics that is equitable for *all students and subgroups of students*.

There are several features of the Smarter Balanced summative assessments that support equitable assessment across all groups of students. For example, the assessments are developed using the principles of universal test design; test accommodations are provided for students with disabilities; and Spanish-language versions of the math assessments will be developed. In addition, there is a specific work group for accessibility and accommodations, and the Consortium has developed seven sets of guidelines to facilitate accessibility of the assessments. These include general accessibility guidelines for item writing and reviewing (Measured Progress & ETS, 2012) and guidelines for creating audio, sign language, and tactile versions of the items. The Consortium also developed guidelines for item development that aim toward reducing construct-irrelevant language complexities

for English language learners (Young, Pitoniak, King, & Ayad, 2012), and comprehensive guidelines for bias and sensitivity (ETS, 2012b). These documents underscore the Consortium's commitment to fair and equitable assessment for all students, regardless of their sex, cultural heritage, disability status, native language, or other characteristics.

Irrespective of these proactive activities designed to promote equitable assessments, studies must be done to provide validity evidence that the assessments are fair for all groups of students. Many of the equity issues are delineated in the most recent version of the NCLB Peer Review Guidance (U.S. Department of Education, 2009b). For example, these guidelines recommend providing translations in appropriate languages and formats (p. 66), and they require statistical evidence of comparability across different language versions of assessments (p. 36). These guidelines also require that all students be included in the assessment, regardless of disability or English language proficiency status.

Of these requirements, statistical evidence of comparability across the English- and Spanish-language versions of the math assessments, and across standard and accommodated test administrations, is particularly important. For example, the Standards assert, "When multiple language versions of a test are intended to be comparable, test developers should report evidence of test comparability" (AERA et al., 1999, p. 99). Similarly, the ITC's Guidelines on Test Adaptation (Hambleton, 2005) state that "Test developers/publishers should apply appropriate statistical techniques to (a) establish the equivalence of the language versions of the test, and (b) identify problematic components or aspects of the test that may be inadequate in one or more of the intended populations" (p. 22). Thus, empirical analyses to evaluate the comparability of the English- and Spanish-language versions of the math summative assessments are needed. Similar evidence will be needed to evaluate the comparability of standard and accommodated tests.

To evaluate the degree to which the summative assessments are fulfilling the purpose of providing valid, reliable, and fair information that is equitable for all students, several studies are recommended. These studies are categorized here as validity evidence based on all five sources of evidence listed in the Standards.

Validity Studies Based on Test Content. Validity studies based on test content to support the equitability of the assessments will be based on the degree to which the planned universal test design, guidelines for assessing English language learners, and other fairness guidelines are implemented and followed. Documents regarding sensitivity review, and how items that were flagged for DIF were handled, should be reviewed. The test development processes and scoring processes are designed to minimize sources of construct-irrelevant variance that would inhibit fairness. The degree to which these procedures are followed and documented should be audited. Part of this audit should ascertain the degree to which students with disabilities, underrepresented minorities, and English language learners were included in the field tests, and the degree to which their special characteristics were addressed in scoring.

Validity Studies Based on Internal Structure. When evaluating the comparability of different variations of a test, such as different language versions of an assessment or accommodated test administrations, validity studies based on internal structure are most common (Sireci, Han, & Wells, 2008). These studies most often involve multi-group confirmatory factor analysis (CFA) (e.g., Ercikan & Koh, 2005). Weighted (multi-group) multidimensional scaling (MDS) has also been used for this purpose (e.g., Robin, Sireci, & Hambleton, 2003; Sireci & Wells, 2010). Both CFA and MDS involve simultaneous analysis of the dimensions underlying an assessment, and are used to assess whether the dimensionality is invariant across different versions of an exam. The CFA approach allows for statistical tests of different levels of invariance (number of dimensions, item factor loadings, correlations among factors, errors associated with factor loadings). The MDS approach does not typically involve statistical tests of invariance, but because it is exploratory, the dimensionality does not need to be modeled a priori.

Multi-group analyses of dimensionality can also be used to evaluate the comparability of scores for different subgroups of students who take the same test. For example, Day and Rounds (1998) used weighted MDS to look at structural invariance of an assessment across ethnic groups, and Marsh, Martin, and Jackson (2010) used multi-group CFA for this same purpose. The validity research agenda should use multi-group CFA or MDS to evaluate the invariance of test structure across diverse groups of students taking the standard versions of the summative assessments, as well as across students taking the standard and accommodated versions of the assessments.

In addition to comparing the dimensionality of the summative assessments across diverse groups of students, simpler analyses based on internal structure should also be performed. Essentially, these analyses involve breaking down the results of all studies of measurement precision to the subgroup level. Reliability estimates, conditional standard error functions, DC and DA estimates, and average standard errors should be reported for all subgroups and all different versions of the assessments. Given that reliability estimates are influenced by variability in students' responses, comparisons of measurement precision are better if based on estimates of the standard error of measurement.

One other important source of validity evidence to support equitable assessment for all is analysis of DIF across test variations and across subgroups of students. There are numerous procedures for evaluating items for DIF, and because excellent descriptions of these procedures exist (e.g., Clauser & Mazor, 1998; Holland & Wainer, 1993), they are not described here. DIF studies conducted for the summative assessments should include an effect size criterion to distinguish statistically significant DIF from substantively meaningful DIF (i.e., reflect construct-irrelevant variance). The presence of DIF does not necessarily indicate bias, and so DIF studies must be followed up by qualitative analysis to try to interpret the source of DIF. Finally, the DIF studies should evaluate the *aggregate* effect of DIF at the total test score level, or at least estimate how the presence of some DIF items may affect the typical test taker from a subgroup.

Validity Studies Based on Response Processes. The studies involving validity evidence based on response processes for purpose 1 are relevant here in that relevant subgroups of students should be included in those studies and the results should be broken down by subgroup. In particular, the amount of time that different groups of students take to respond to items, both with and without accommodations, should be studied. Any cognitive interviews or think-aloud protocols that are conducted to evaluate the skills measured by items should be inclusive in recruiting students. In addition, specific studies to evaluate accommodations for English language learners or students with disabilities should be conducted to determine whether the students are using the accommodations and find them helpful (e.g., Duncan et al., 2005).

Validity Studies Based on Relations to Other Variables. Two types of studies based on relations to other variables are relevant for validating that the summative assessments are equitable for all subgroups of students. The first are differential predictive validity studies that evaluate the consistency of the degree to which the assessments predict external criteria across subgroups of students. Zwick and Schlemer (2004) provide an excellent example of this type of analysis with respect to the differential predictive validity of the SAT across native English speakers and non-native English speakers. These studies will be particularly relevant for the "on track" and "college and career readiness" standards associated with the summative assessments. Of course, the caveats that were mentioned earlier regarding the validity of the external criteria apply here.

The second type of study involves a grouping variable as the external variable. Experimental studies that have looked at test accommodations fall into this category. For example, in some studies, students with and without disabilities are randomly assigned to test accommodation or standard test administration conditions. The validity hypothesis investigated is one of "differential boost," which states that students with disabilities will have larger score differences across the accommodated and standard conditions than students without disabilities, and that their scores will be higher in the accommodated condition (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000).

Non-experimental studies using grouping variables could also be conducted using an expected hypothesis of no difference across groups. For example, using changes in students' scale scores over time as the dependent variable, comparisons could be made across students of different ethnic groups, SES, sexes, and other demographic characteristics.

In addition to the studies previously described in this section, all other studies conducted on the general population could be broken down by subgroup to evaluate consistency of the results across subgroups, where sample sizes permit. For example, if multitrait-multimethod studies are conducted, a study of the invariance of results across subgroups may prove interesting.

Validity Studies Based on Testing Consequences. The analysis of the results from the summative assessments across subgroups of students will be a good starting point for understanding if there are differential consequences for certain types of students. In describing validity studies based on testing consequences for other purposes of the summative assessments, we discussed investigating the effects on instruction, teacher morale, and students' emotions and behaviors (e.g., dropout, course-taking patterns). These results should also be broken out by subgroup, but more importantly, the changes in instructional decisions for students should be investigated at the subgroup level. Important analysis questions include: Are minority students dropping out of school at higher rates than non-minorities? Are the success rates for remedial programs higher for certain types of students?

VI. Validity Agenda for Interim Assessments

The Smarter Balanced interim assessments differ from the summative assessments in that they are optional, include both secure and non-secure components, are customizable across users, 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 CCSS. The validity studies described for the summative assessments are essentially all relevant to the interim assessments, but additional validation work needs to address the degree to which the interim assessments provide the intended diagnostic information and are useful to teachers, administrators, and other educators for improving instruction and student learning.

As indicated in Chapter III, four purpose statements for validation are associated with the interim assessments. The proposed studies to support the validity of these statements are described in this section.

Interim Assessment Purpose 1:

Provide valid, reliable, and fair information about students' progress toward mastery of the skills measured in ELA and mathematics by the 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 interim assessment scores are on the same scale as those from the summative assessments. As indicated in Table 2 (p. 15), the studies providing this evidence will primarily be based on test content, internal structure, and response processes.

Validity Studies Based on Test Content. The content validity studies described for the summative assessments will gather data relevant to the interim assessments. However, an additional level of analysis will be required to support the validity of reporting students' performance at the content cluster levels. The sample results of a summary of a content validity study that were reported in Figure 4 (p. 30) suggest how results could be summarized for the content clusters targeted by the interim assessments. Moreover, the data from such studies could be used to select the best items for interim assessment purposes. That is, items that are rated as measuring their intended CCSS "very well" could be selected for the interim assessment item bank.

The interim assessments are intended to help teachers focus assessment on the most relevant aspects of their instruction at a particular point in time. Thus, the interim assessments should better align with teachers' instruction, if the content clusters are appropriately selected. To evaluate this intended benefit of the interim assessments, surveys could be given to teachers regarding the instructional objectives that they cover at several points during the school year (i.e., scope and sequence survey). Then, the content clusters that were administered to these teachers' students at specific points in time can be evaluated ex post facto, and the match between what was taught and what was assessed can be calculated. This type of survey could be coupled with survey questions regarding the utility of the interim assessments, which is relevant to purpose 2.

Validity Studies Based on Internal Structure. Scores from the comprehensive interim assessments are intended to be on the same scale as those from the summative assessments, to best measure students' progress toward mastery of the knowledge and skills measured on those assessments. This intent requires linking the scores from the interim and summative assessments. Given that many of the items in the interim assessment item bank will also be used on the summative assessments, it is assumed that some type of common item equating will be used to place students' performance on the interim assessments on the summative assessment score scale. This equating should be evaluated to support the inferences about how well students are likely to do on the summative assessments based on their interim assessment scores. Studies in this area would

include an audit of the equating procedures, such as analysis of equating error and analysis of DIF of equating items across groups of students defined by state, ethnicity, or other factors (or a more formal population invariance study; Dorans, 2004). In addition, the degree to which interim assessment items fit the IRT models determined by the summative assessment scale should be ascertained. The fit of the equating items to this model will be of particular interest.

Also under the realm of internal structure is evidence regarding the reliability or measurement precision of scores from the interim assessments. Less measurement precision relative to that of the summative assessments is tolerable because (a) the stakes are lower, (b) there will be multiple assessments, and (c) these assessments supplement the summative assessments, on which higher-stakes decisions are based. However, studies should be conducted to ascertain the reliabilities and errors of measurement associated with any scores reported from the interim assessments so that they can be properly interpreted. If achievement level classifications are made on the basis of these assessments, then estimates of DC and DA should also be calculated.

Studies should also be conducted to evaluate the quality and accuracy of local scoring of the performance tasks associated with the interim assessments. Having trained scorers rescore samples of locally scored tasks, and the degree to which local scorers can assign similar scores to training sets of responses, will provide evidence regarding the quality of local scoring.

Validity Studies Based on Response Processes. Interim Assessment Purpose 1 relates to skills measured on the summative assessments, and so the validity studies based on response processes that were described for the summative assessments are relevant here in order to confirm that the items are measuring higher-order skills. The response process studies for Summative Assessment Purpose 1 should include items that will be used on the interim assessment. The results from these studies should be used to “assure that each item or task clearly elicits student responses that support the relevant evidence statements and thus are aligned to the associated claims and standards” (ETS, 2012c, p. 4).

Interim Assessment Purpose 2:

Provide valid, reliable, and fair information about students’ performance at the content cluster level, so that teachers and administrators can track student progress throughout the year and adjust instruction accordingly

As shown in Table 2, validity evidence to support this purpose of the interim assessments will rely on studies of test content, internal structure, and testing consequences.

Validity Studies Based on Test Content. Assuming that the content validity/alignment studies described for the summative assessments are conducted, all items on those assessments will be rated regarding the degree to which they measure their intended CCSS and their intended cognitive skills. These studies should be extended to include the items on the interim assessments that do not overlap with the summative assessments. However, an additional study is needed to support purpose 2. A study should be conducted to confirm that the content clusters associated with the interim assessments represent helpful groupings of CCSS that are useful for tracking progress and adjusting instruction. These studies would evaluate whether the specific groupings of standards from the CCSS into content clusters is instructionally beneficial.

Like all content validity studies, this study would require SMEs. Rather than reviewing items, the SMEs would review the CCSS that were used to create the content clusters for each claim area. Their task could be to group the standards in a way that would be best for providing instructionally relevant information. Their groupings of standards could then be compared to how the standards were grouped into the content clusters, and the consistency across the actual and SME-derived clusters could be calculated. Alternatively, the SMEs could review the content clusters and rate them

for their instructional relevance, and make comments about whether and how they might be rearranged.

Validity Studies Based on Internal Structure. Information regarding the reliability and measurement error of cluster-level score reporting should be provided. In addition, the degree to which different clusters are correlated should also be reported, to see if clusters measuring different assessment targets or claims correlated less than clusters measuring the same claims and targets. A multitrait-multimethod approach could be used, using the different item formats and different claim areas as methods and traits, respectively (Pitoniak, Sireci, & Luecht, 2002).

Validity Studies Based on Testing Consequences. The interim assessments are designed to “provide more immediately actionable data for teachers and students” (ETS, 2012c). A primary validity question to be studied is: Do the content cluster results help teachers and administrators track student progress and adjust instruction? To assess the effects on instruction, studies should be conducted to (a) track the use of the interim assessments and their associated supports (e.g., user tutorials), (b) assess the degree to which teachers and administrators find the system easy to navigate, and (c) assess the degree to which teachers and administrators value the information provided and use it to adjust instruction. Studies could also be conducted to ascertain students’ impressions of the system.

Tracking the use of the interim assessments should be straightforward, assuming that most of the assessments are accessed online and that these testing occasions are captured by the system. Procedures should be in place to track any uses that are not online. Surveys of teachers and administrators will be needed in order to understand the degree to which these educators find the system useful and easy to navigate. Surveys of teachers and administrators will also be needed to ascertain the effects on instruction. As part of that study, “high use” teachers and schools should be identified and selected for further inquiry. Surveys, interviews, and focus groups of these teachers should be conducted, to learn about how they used interim assessment results to improve instruction.

Interim Assessment Purpose 3:

Provide valid, reliable, and fair information about individual and group (e.g., school, district) performance at the claim level in ELA and mathematics, to determine whether teaching and learning are on target.

As shown in Table 2, validity evidence to support this purpose of the interim assessments will rely on studies of internal structure, relations to other variables, and testing consequences.

Validity Studies Based on Internal Structure. This purpose statement is similar to purpose 2, with the difference being that rather than a focus at the content cluster level, the focus here is on the claim level. The studies described for purpose 2 are all relevant here. The additional studies needed would need to evaluate the reliability and precision of the claim scores at the group level. It is assumed that claim-level information will be provided by the interim assessments during the school year, and so estimates of the precision of this information should be provided, using the same types of internal structure studies described for purposes 1 and 2.

Validity Studies Based on Relations to Other Variables. Given that the interim assessments will provide information at the claim level throughout the school year, it would be good to study the degree to which the information provided for individual students or groups of students is consistent with other measures of their performance relative to the CCSS. One way to study this relationship is to see how well the claim scores for the interim assessments predict claim scores on the summative assessments. In particular, it would be interesting to assess the degree to which students who are considered “on target” or “not on target” are classified similarly on the summative assessments. More interesting, however, would be to qualitatively study students who are mispredicted. That is, if

a student did poorly on an interim assessment but well on a summative assessment, is that a success story or a story of poor measurement by the interim assessment? If other measures of student achievement are available, they would be helpful for shedding light on this issue, but it may be difficult to find other measures tied to the same CCSS that specific interim assessments are measuring. Nevertheless, assessments such as NWEA's Measures of Academic Progress or Curriculum Associates' iReady assessment may be relevant.

Validity Studies Based on Testing Consequences. As mentioned for purpose 2, the intended consequence of the interim assessments is to connect the assessments to instruction to improve student learning. The validity studies based on testing consequences that were described for purpose 2 are all relevant here, with the only difference being that the information provided would be at the claim level and would be extended to groups of students. Therefore, the studies described earlier should include these factors to provide validity evidence in support of purpose 3. In addition, should in-class activities (classroom interaction tasks) become part of the interim assessment system, their effectiveness should be a focus of the surveys, interviews, and focus groups associated with the studies mentioned earlier.

Interim Assessment Purpose 4:

Provide valid, reliable, and fair information about student progress toward the mastery of skills measured in ELA and mathematics *across all students and subgroups of students.*

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 purposes 1 and 2 provide the starting point for equitable measurement across all students. The validity studies based on internal structure should report any estimates of reliability, measurement precision, DC, or DA separately for all subgroups of students, and for students who take different variations of the interim assessments. In addition, it should be documented that access to the interim assessments has been provided to all students, as was discussed in relation to the summative assessments. Such access should include appropriate test accommodations for students with disabilities and English language learners.

The Peer Review Guidance for NCLB assessments stipulates that states should "Provide written documentation of criteria for local assessments, which ensures technical quality and comparability to state assessments of locally used tests for ALL subgroups and content areas (includes modified/alternate assessments)" (U.S. Department of Education, 2009b, p. 32). The interim assessment system allows states and districts to create their own assessments from the banks of items, and so the technical quality of these local assessments will need to be studied to ensure that they provide comparable measurement across all groups of students.

VII. Research Agenda for Formative Assessment Resources

The third component of the Smarter Balanced Assessment Consortium is *formative tools and processes*, referred to in this report as *formative assessment resources*. These resources are not assessments per se, and so their evaluation does not neatly fit into the Standards' five sources of validity evidence. Rather, these resources are intended to work with the summative and interim assessments to increase their utility for improving instruction and helping students learn. Essentially, the formative assessment resources are what puts the "balance" in the Smarter Balanced Assessment Consortium.

The purposes of the formative assessment resources that are the focus of the comprehensive research agenda were listed in Chapter III, and, for convenience, are repeated here.

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. Help teachers and other educators use 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.

To accomplish these goals, the formative assessment resources will provide tools and professional development materials including a "Digital Library," learning modules (lesson plans, templates, curriculum resources, evidence collection tools, video clips of classroom instruction and teacher analysis, descriptive feedback strategies, follow-up planning materials), online assessment literacy training products, webinars, tutorials, and PowerPoint presentations. To oversee the development, implementation, and maintenance of these resources, extensive collaboratives will be established, including:

- National Advisory Panel
- Digital Library Review Board
- State Leadership Teams
- State Networks of Educators
- Formative Assessment Practices and Professional Learning Work Group

The research agenda for this component of the Consortium will be an evaluation of the products developed for these purposes and of the processes for developing them. Studies comprising this evaluation should involve (a) confirming the development and successful implementation of all planned formative assessment resources; (b) evaluating usage statistics of all tools and other resources; (c) review of all documents supporting the system; (d) comprehensive surveys of the collaborative leadership involved in overseeing the products and processes; (e) comprehensive surveys of users of the resources (teachers, administrators, students, parents); and (f) case studies of teachers and administrators who are frequent users of the resources. It should also be confirmed that teachers were involved in the development and review of these materials.

Confirming Development and Successful Implementation of Products

The RFP for the “Digital Library with Formative Assessment Practices and Professional Learning Resources for Educators,” hereafter referred to as RFP-23, specifies the development of several products using specific processes. An important step in the evaluation of the formative assessment resources is to confirm that all of the deliverables associated with this contract were satisfied. For example, RFP-23 calls for the development of at least 50 exemplar instructional modules (p. 26). The successful creation of these modules, and other tasks, will be audited as part of the evaluation. In addition, goals related to the review and implementation of all resources will be reviewed in this evaluation. This step will merely confirm that the intended products and activities occurred and note the timeliness of the deliverables. The quality of the products and their implementation will be evaluated using other activities described later in this chapter.

Evaluating Usage Statistics

The formative assessment resources are designed to be used by teachers, administrators, and even parents and students. If these resources are not understood and found useful, the system will be unbalanced, which will inhibit the goals of the entire Consortium. One way to evaluate the utility of the resources is to analyze their usage statistics. RFP-23 specifies reporting monthly usage statistics (p. 71). These statistics should be analyzed over time. Formative evaluation should inform the Smarter Balanced leadership about which resources are being used and which are not, so that better advertising or improvement of the underutilized resources can be considered. Analysis of usage data should be broken down by state, and by important subcategories within states, such as type of school, geographic region, percentage of certain subgroups of students within a school (English language learners, low-SES, etc.), and, where possible, demographics of the users.

Document Review

RFP-23 specifies several documents that are important to the integrity of the formative assessment resources. These documents include:

- Comprehensive development strategy
- Biannual implementation reports
- Documentation of component plans and processes
- Description of recruiting and creation of leadership committees (State Leadership Teams, State Networks of Educators)
- Records of decision-making by leadership committees
- Technical documentation of system components

These documents will be reviewed to ensure that products are developed as intended and processes are followed. Any problems discovered in the documents should be followed up on to see if they were properly resolved. In addition, RFP-23 requires the contractor to perform and document quality assurance testing (pp. 69–70). This documentation will also be reviewed as part of the evaluation. Monitoring reports on user comments (p. 71) will also be reviewed and reported on.

Surveys, Interviews, and Focus Groups of Leadership

The plan for developing, implementing, and improving the formative assessment resources calls for full participation of educators throughout the Consortium. In particular, the State Networks of Educators will involve carefully selected end-users of the resources. In the evaluation, the five aforementioned collaboratives of leaders (National Advisory Panel, Digital Library Review Board, State Leadership Teams, State Networks of Educators, Formative Assessment Practices and

Professional Learning Group) will be solicited to participate in surveys, interviews, or focus groups to obtain their impressions of the process, the quality of the products, and the degree to which the formative assessment resources are accomplishing the intended goals. In addition, the intended representation of the membership of these committees with respect to geographic region, subject expertise, representation of special populations, and other characteristics will be evaluated.

Surveys of Users

The evaluation activities previously described will provide information on the quality of the products and processes and the degree to which users are accessing the resources. However, it is also critical to gather information regarding the degree to which the resources are perceived as being helpful to educators. RFP-23 includes the development of a survey to assess the effectiveness of the regional meetings (p. 23). The results from that survey should be considered in the evaluation. More importantly, however, we recommend that the research agenda include large-scale surveys of all users. Given that the bulk of the resources must be accessed online, *we recommend that user surveys be implemented as part of the system*. That is, at strategic points in time, users should be required, or heavily encouraged, to take brief surveys, for the Consortium to obtain their opinions regarding the usefulness of the materials and how they use the resources in their instructional practices. The surveys should target the specific aspects of the resources (e.g., lesson plans, evidence collection tools, assessment literacy training products, understanding how to use summative and interim data to improve instruction, etc.). Surveys to evaluate training programs delivered as part of the implementation of the resources (e.g. RFP-23, p. 65) are also needed. These surveys are needed in order to provide evidence that the formative assessment resources are having an impact on classroom practices.

Teacher survey data could also be used to create an implementation index for participating teachers, and those data could be correlated with students' test scores. In particular, it would be interesting to correlate teachers' implementation data with the progress that students make *within the school year* while they have the teacher. If all aspects of the system work as intended, teachers who successfully use the formative assessment resources will be able to use the summative and interim assessment results to improve instruction, and will see greater gains for their students, relative to comparable teachers who do not use the resources.

It is also important to gather data on the degree to which parents, students, teachers, administrators, and others understand the reports from the summative and interim assessments. These data can be gathered using surveys to obtain opinions of the reports, and also by testing these individuals regarding the accuracy of their interpretations (Wainer et al., 1999).

Case Studies of Frequent Users

The usage data for the formative assessment resources can be used to identify teachers and administrators who are frequent users. A sample of these frequent users can be selected and recruited for in-depth study of how they use the resources. The appropriateness of their practices can be documented, and ideas for improving the resources, and for sharing the lessons learned by these teachers and administrators, can be reported.

VIII. Summary: The Smarter Balanced Assessment Consortium Validity Argument

The preceding chapters describe a multitude of studies that comprise the comprehensive research agenda for the Smarter Balanced Assessment Consortium. The presentation of the agenda according to the different components of the system may result in two misleading perceptions. These potential misleading perceptions are:

- The research agenda is too ideal to be practical because the agenda is too voluminous and optimistic.
- The research agenda is fragmented and so does not address the holistic goals of the Consortium.

In this chapter, we put those potential misperceptions to rest by illustrating the integration of studies across the various components and illustrating how many of the studies are already addressed in the test development and formative assessment resources development activities.

The integration of the various studies results in an agenda that, if properly implemented, can provide a convincing validity argument to support the goals of the Consortium as stated in its Theory of Action (Appendix A). Bennett (2010) posited six questions that should be posed to evaluate a theory of action for a comprehensive assessment system such as Smarter Balanced. These seven questions are:

- Is the theory of action logical, coherent, and scientifically defensible?
- Was the assessment system implemented as designed?
- Were the interpretive claims empirically supported?
- Were the intended effects on individuals and institutions achieved, and did the postulated mechanisms appear to cause those effects?
- What important unintended effects appear to have occurred? (p. 82)

The first question can be addressed by a thoughtful review of the Smarter Balanced Theory of Action as a preliminary step in the evaluation. Our impression is that the theory is defensible, which is supported by the fact that we were able to create a comprehensive research agenda to address its goals. The second question can be answered by analysis of the results from the studies outlined in this report, specifically the audit studies listed in Chapters III and VII and the studies regarding validity evidence based on testing consequences that involve surveys, interviews, and focus groups of stakeholders (described in Chapters IV through VII).

What most people think about when considering validation of an assessment system are the third and fourth questions posed by Bennett (2010). We, and many others (e.g., Haertel, 1999; Messick, 1989; Shepard, 1993), would also include the sixth question. These three questions require validity evidence beyond typical test development activities, and require evidence stemming from all five sources stipulated in the Standards. It is around these three questions that the majority of studies described in Chapters V through VII are centered.

The Smarter Balanced Theory of Action is based on seven principles (Smarter Balanced, 2010). These principles are presented in Appendix A and are presented here in more abbreviated form:

1. Assessments are grounded in a thoughtful, standards-based curriculum and are managed as part of an integrated system.
2. Assessments produce evidence of student performance.
3. Teachers are integrally involved in the development and scoring of assessments.
4. The development and implementation of the assessment system is a state-led effort with a transparent and inclusive governance structure.

5. Assessments are structured to continuously improve teaching and learning.
6. Assessment, reporting, and accountability systems provide *useful information on multiple measures* that is educative for all stakeholders.
7. Design and implementation strategies adhere to established professional standards. (pp. 32–33)

A review of the purpose statements on which this comprehensive research agenda is based (see Chapter III) makes clear that the agenda is focused on evaluating the degree to which these principles are realized. To pull the comprehensive research together—that is, to document the validity argument for Smarter Balanced in a coherent manner to best inform stakeholders and the general public—a report should be produced that indicates how the various pieces of evidence gathered through the research agenda confirm that these seven principles are realized. If the research agenda outlined in this report is followed, it will provide ample evidence that could be organized in a reader-friendly report that is organized around these seven principles. It is clear that the research agenda outlined here addresses the seventh principle. Our review of Smarter Balanced activities to date supports the fourth principle, and evidence for the collaboration could easily be documented. The remaining five principles would be supported by evidence from the studies described in this report.

Summarizing the Validity Evidence

As promised earlier in this chapter, the validity studies described in this report will appear less daunting when the overlap of studies across the different purposes and components of the Smarter Balanced assessment system is accounted for. This integration is presented in Tables 7 and 8. Table 7 presents brief descriptions of each proposed study in the form of short labels, indicates the purposes that each study addresses, and provides a unique number for each study. It also lists the page numbers in this document that refer to each study. Table 8 uses this numbering system to illustrate the places where such studies are already accounted for in current or planned Smarter Balanced activities. Table 8 is also available as an Excel file, so that its data can be sorted by columns to facilitate different research planning activities. It may be tempting to prioritize the studies based on the number of check marks in each row of Table 7, but because the purposes in the columns are not equal in importance, and because the contribution of each study to the validity argument will not be equal, such an interpretation would be an oversimplification.

Table 7. Listing of Studies by Source of Evidence and Testing Purpose.

Study Number and Description	Page Numbers	Evidence Sources	Summative Assessment Purpose							Interim Assessment Purpose				Formative Resources Purpose				
			1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5
1 Audit of test construction practices	16–18	1, 3	√	√	√	√	√	√	√	√	√	√	√			√	√	√
2 Analysis of measurement precision	17–18, 34, 36–37, 51–52, 53, 56, 58–59	3		√	√	√	√	√	√									
3 Audit of test administration	17, 18	1, 5	√						√				√		√			
4 Evaluation of scoring	17, 19	1, 3	√	√	√	√		√	√	√	√	√	√	√	√	√	√	√
5 Analysis of scaling and equating	17, 19, 46–48, 58–59	3		√	√			√	√	√								
6 Evaluation of standard setting	17, 19–21, 36–45	1, 3, 4	√	√	√	√	√	√	√			√		√			√	
7 Evaluation of fairness	17, 22, 52, 59	1–5	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
8 Evaluation of equitable particip. & access	17, 23–24, 52, 54, 59	1, 5	√						√				√				√	√
9 Audit of test security	17, 24–25, 48	3, 4	√	√	√			√		√								
10 Content validity and alignment	25–31, 36, 39–40, 46, 53, 56–58, 86	1	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
11 Evaluating ECD	25, 31	1, 2	√	√	√	√				√	√	√						
12 IRT residual analysis	31–34, 46	3		√		√		√	√	√								
13 Reliability and standard error estimation	17–19, 31–34, 36–37, 50–52, 53–54, 56–59	3	√	√	√	√	√	√	√	√	√							
14 Cognitive skills and item response time	24, 35, 54, 56	2	√				√		√		√							

Study Number and Description	Page Numbers	Evidence Sources	Summative Assessment Purpose							Interim Assessment Purpose				Formative Resources Purpose				
			1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5
15 Cognitive interviews, think-aloud	35, 54, 56	2	√				√		√		√							
16 Decision consistency and accuracy	36–38,41, 56–57, 59	3		√	√	√		√	√	√		√						
17 Cut-score standard errors	36–37	3		√	√	√		√	√	√		√						
18 Criterion-related validation of “on track”	37–38	4		√														
19 Educator interviews, focus groups, surveys	38–39, 44–45, 49–50, 58–59	5		√	√													
20 Criterion-related validation of readiness	39–45	4		√	√	√		√		√								
21 Surveys of postsecondary educators	45, 62	5			√	√												
22 Analysis of enrollment, dropout, courses	38, 45, 55	5					√		√				√					
23 Teacher morale surveys	45, 52, 55	5					√	√		√	√	√	√	√		√	√	√
24 Teacher surveys on changes in students	45, 49, 52–53, 55, 59, 61–62	5		√	√	√	√		√	√	√	√	√	√	√			√
25 Student morale and aspirations surveys	45, 52	5			√			√										
26 Evaluation of vertical scale	46–48	3		√	√	√		√	√	√								
27 Criterion-related studies re: gain/growth	49, 52, 59	4		√		√		√	√	√	√	√						
28 Follow-up on specific student decisions	49–50	5		√	√	√	√	√	√	√	√	√	√	√	√		√	√
29 Sensitivity to instruction	49–50	4	√	√	√	√	√	√	√	√	√	√	√					

Study Number and Description	Page Numbers	Evidence Sources	Summative Assessment Purpose							Interim Assessment Purpose				Formative Resources Purpose				
			1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5
30 Analysis of classroom artifacts	49–50	4, 5					√				√			√	√	√	√	√
31 Score report utility and clarity	31, 49–50, 61–62	5	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
32 Analysis of report usage rates	49–50	5					√							√	√	√	√	√
33 Analysis of reliability of aggregate stats	50–52, 59	3	√					√	√			√		√				
34 Generalizability studies	17–20, 51	3		√	√	√		√										
35 Item parameter drift	48	3		√	√	√				√								
36 Audit of UTD and sensitivity review	54	1	√					√	√				√					
37 Audit of test accommodations	53	1, 5	√					√	√				√					
38 Differential item functioning	17, 18, 21–22, 53–54, 56–57	1, 3	√					√	√				√					
39 Differential predictive validity	54	4	√					√	√				√					
40 Invariance of test structure	19, 53–54, 56–57	3	√					√	√				√					
41 Analysis of group differences	54–55	4		√	√	√		√	√				√					
42 Multitrait-multimethod	55, 57	3, 4	√						√		√	√	√					
43 Scope and sequence curriculum survey	56–57	1, 5					√				√	√		√	√	√	√	√
44 Validation of content clusters	56	1, 3									√	√						
45 Analysis of interim usage statistics	57–58	5								√	√	√	√					

Study Number and Description	Page Numbers	Evidence Sources	Summative Assessment Purpose							Interim Assessment Purpose				Formative Resources Purpose				
			1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5
46 Surveys, interviews, focus groups of (high) users of interim assessments	57–58	5								√	√	√	√					
47 Audit of formative resources development and implementation	61–62	1, 5												√	√	√	√	√
48 Analysis of usage stats for formative	61–62	5												√	√	√	√	√
49 Surveys of collaborative leadership	61–62	5												√		√	√	√
50 Educator formative assessment surveys	61–63	5												√	√	√	√	√
51 Formative assessment user surveys	62	5												√	√	√	√	√
52 Parent, student formative surveys	48–49, 63	5												√				√
53 Case studies of frequent users	62	5												√	√	√	√	√
54 Critique of Theory of Action	63–64	5	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
55 Summary of validity evidence acc. to 7 principles	64–68	5	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

Note: Evidence Sources: 1 = Test Content, 2 = Response Processes, 3 = Internal Structure, 4 = Relations to Other Variables, 5 = Testing Consequences

Table 8 (to be populated): Connecting Recommended Studies to Current Activities and RFPs

Study and Number	Source of Evidence	Contract	Summative Assessments	Interim Assessments	Formative Assessment Resources
1 TC audit					
2 Meas. precision					
3 Administration audit					
4 Evaluation of scoring					
5 Scaling and equating					
6 Standard setting					
7 Evaluation of fairness					
8 Equity					
9 Audit of test security					
10 Content validity					
11 Evaluating ECD					
12 IRT residual analysis					
13 Reliability and SE					
14 item response time					
15 Cognitive interviews					
16 DC, DA					
17 Cut-score SE					
18 Criterion-related OT					
19 Educator surveys					
20 Readiness					
21 Postsecondary surveys					
22 Dropout					
23 Teacher morale					
24 Change surveys					

Study and Number	Source of Evidence	Contract	Summative Assessments	Interim Assessments	Formative Assessment Resources
25 Student morale					
26 Vertical scale					
27 Gain (growth)					
28 Student decisions					
29 Sensitivity					
30 Classroom artifacts					
31 Score reports					
32 Report usage rates					
33 Aggregate stats					
34 G-studies					
35 Item parameter drift					
36 UTD and sensitivity					
37 Test accommodations					
38 DIF					
39 Diff. prediction					
40 Invariance					
41 Group differences					
42 MTMM					
43 Scope and sequence					
44 Content clusters					
45 Interim usage					
46 Surveys high users					
47 Formative audit					
48 Formative usage					
49 Collabor. leadership					

Study and Number	Source of Evidence	Contract	Summative Assessments	Interim Assessments	Formative Assessment Resources
50 Educator FA surveys					
51 FA user surveys					
52 Parent/student surveys					
53 Case studies: users					
54 Theory of Action					
55 Summary of validity					

IX. Ongoing Validation Activities and Support Systems

Validation can be thought of as a great job for a masochist because, in a sense, one can never absolutely “prove” that an assessment is totally valid for the complex purposes to which it is put (Haertel, 1999), and because assessments are dynamic, and they, and the populations that they assess, change over time, validation is an ongoing, essentially perpetual, endeavor. Nonetheless, at some point, decisions must be made regarding whether sufficient evidence exists to justify the use of a test for a particular purpose. Most of this report has focused on the purpose of conducting studies to provide such evidence and documenting the evidence into a coherent validity argument that would satisfy professional testing standards, federal peer review, and legal challenges. However, our professional responsibilities also require us to think toward the future, beyond the current funding for Smarter Balanced, and consider the potential positive and negative consequences that should be addressed in longer-range validation studies.

At this juncture, a few potential validity activities appear in the crystal ball. One is studying the degree to which products and processes provided by the Consortium persevere and are used over time. The Consortium’s processes, products, and activities are designed to produce an enduring collaboration and resources that should outlive the Consortium. Thus, studying the long-term effects of Smarter Balanced on instruction, within and outside the Consortium states, would be an interesting research area.

Another area of interest is the specific uses of the Smarter Balanced assessments and formative resources beyond the currently anticipated uses. It is quite possible that states, districts, and schools will use the assessments for purposes that they think are useful and valid, but that are not currently anticipated. Some of these uses may be appropriate and creative; others may be problematic or even damaging. States and districts will certainly use some assessments and tools for educator accountability, and so the validity of such use is an area in need of future research.

Although all important areas of future research cannot be anticipated at this time, it is still wise to consider the support systems that Smarter Balanced can put in place to facilitate future validity research. For example, other large-scale assessment programs, such as NAEP, TIMSS, and PISA, make data available for secondary analyses. Occasionally, these programs provide grant money to support such secondary analyses. The types of studies to be funded can be specified in advance, or, preferably, applicants for funding could be asked to submit their own ideas for research to study what they believe are important validity questions.

Another example of a support system is the College Board’s “validity research study service.” This service is essentially a data-sharing agreement between the College Board and postsecondary institutions, whereby the institutions can send course grade information to the College Board and it will match the data with SAT scores and other College Board assessment scores. These matched data sets can then be used to conduct local validity studies for each institution.

In considering potential validity studies that will be important in the future, and by establishing research support systems, validity research for Smarter Balanced can outlive the formal research studies that will comprise the documented validity argument for the Consortium.

References

- Achieve, Inc. (2006). *An alignment analysis of Washington state's college readiness mathematics standards with various local placement tests*. Cambridge, MA: Author.
- ACT. (2005a). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Process report*. Iowa City, IA: Author.
- ACT. (2005b). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Special studies report*. Iowa City, IA: Author.
- ACT. (2005c). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Technical report*. Iowa City, IA: Author.
- ACT. (2006). *Ready for college, ready for work. Same or different?* Iowa City, IA: Author.
- ACT. (2010). *Issues in college readiness: What are ACT's college readiness benchmarks?* Iowa City, IA: Author.
- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40, 955–959.
- Allen, J., & Sconing, J. (2005). *Using ACT assessment scores to set benchmarks for college readiness* (ACT Research Report Series 2005-3). Iowa City, IA: ACT.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Evaluation Association. (2004). *Guiding principles for evaluators*. Retrieved from <http://www.eval.org/publications/guidingprinciples.asp>
- Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, 8, 70–91.
- 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(3), 21–29.
- Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement*, 12, 261–280.
- Brennan, R. L. (2002). *Estimated standard error of a mean when there are only two observations* (CASMA Technical Note Number 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.
- Brennan, R. L. (2011). *Using generalizability theory to address reliability issues for PARCC assessments: A white paper*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

- Briggs, D. C. (2012, April). *Making inferences about growth and value-added: Design issues for the PARCC consortium*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.
- Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204–226.
- Camara, W. (2012, April). *Defining and measuring college and career readiness: Developing performance level descriptors and defining criteria*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC. Retrieved from <http://research.collegeboard.org/sites/default/files/publications/2012/7/presentation-2012-3-developing-performance-level-descriptors-criteria.pdf>
- Camara, W. J. (2013). Defining and measuring college and career readiness: A validation framework. *Educational Measurement: Issues and Practice*, 32(4), 16–27.
- Camara, W., & Quenemoen, R. (2012). *Defining and measuring college and career readiness and informing the development of performance level descriptors*. Retrieved from <http://www.parcconline.org/sites/parcc/files/PARCC%20CCR%20paper%20v14%201-8-12.pdf>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice*, 17, 31–44.
- Conley, D. T., Drummond, K. V., 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., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, 2, 179–194.
- Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
- Crotts, K., Sireci, S. G., & Zenisky, A. L. (2012). Evaluating the content quality of a multistage-adaptive test. *Journal of Applied Testing Technology*, 13(1), 1–26.
- D'Agostino, J. V., & Bonner, S. M. (2009). High school exit exam scores and university performance. *Educational Assessment*, 14, 25–47.
- D'Agostino, J., Karpinski, A., & Welsh, M. (2011). A method to examine content domain structures. *International Journal of Testing*, 11(4), 295–307.
- D'Agostino, J. V., Welsh, M., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment*, 12, 1–22.
- Davis-Becker, S. L., Buckendahl, C. W., & Gerrow, J. (2011). Evaluating the bookmark standard setting method: The impact of random item ordering. *International Journal of Testing*, 11(1), 24–37.

- Day, S. X., & Rounds, J. (1998). Universality of vocational interest structure among racial and ethnic minorities. *American Psychologist*, 53, 728–736.
- Delisle, D. S. (2012). *Letter to chief state school officers*. Washington, DC: U.S. Department of Education.
- Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, 41, 43–68.
- Duncan, G. D., del Rio Parant, L., Chen, W.-H., Ferrara, S., Johnson, E., Oppler, S., & Shieh, Y.-Y. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Applied Measurement in Education*, 18, 129–161.
- Educational Testing Service (ETS). (2012a). *Smarter09 component 1: Narrative key to understanding the blueprint tables (TAC draft)*. Princeton, NJ: Author.
- Educational Testing Service (ETS). (2012b). *Smarter Balanced Assessment Consortium: Bias and sensitivity guidelines*. Princeton, NJ: Author.
- Educational Testing Service (ETS). (2012c). *Specifications for an interim system of assessment*. Princeton, NJ: Author.
- Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5(1), 23–35.
- Fields, R., & Parsad, B. (2012). *Tests and cut scores used for student placement in postsecondary education: Fall 2011*. Washington, DC: National Assessment Governing Board.
- Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data. *School Psychology Review*, 29, 65–85.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18(4), 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, MD: Johns Hopkins University Press.
- Hambleton, R. K. (1989). Principles and applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York, NY: Macmillan.
- Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting psychological and educational tests for cross-cultural assessment* (pp. 3–38). Hillsdale, NJ: Lawrence Erlbaum.
- Hambleton, R. K., & Han, N. (2004). *Summary of our efforts to compute decision consistency and decision accuracy estimates using IRT item statistics* (Center for Educational Assessment

- Research Report No. 552). Amherst, MA: University of Massachusetts, Center for Educational Assessment.
- Hambleton, R. K., & Rovenelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement*, 10, 287–302.
- Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, & Student Testing.
- Hamilton, L. S. (1994, April). *Validating hands-on science assessments through an investigation of response processes*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of a set of test items. *Applied Psychological Measurement*, 9, 139–164.
- Hill, R. K., & DePascale, C. A. (2002). *Determining the reliability of school scores*. Dover, NH: National Center for the Improvement of Educational Assessment.
- Hill, R. K., & DePascale, C. A. (2003, April). *Adequate yearly progress under NCLB: Reliability considerations*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- International Test Commission (ITC). (2010). *Guidelines for translating and adapting tests*. Retrieved from <http://www.intestcom.org>
- Johnstone, C. J., Altman, J. M., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from <http://www.cehd.umn.edu/nceo/OnlinePubs/StateGuideUD/default.htm>
- Kaira, L. T., & Sireci, S. G. (2010). Evaluating content validity in multistage adaptive testing. *CLEAR Exam Review*, 21(2), 15–23.
- 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.

- Keng, L., Murphy, D., & Gaertner, M. (2012, April). *Supported by data: A comprehensive approach for building empirical evidence for standard setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.
- Kolen, M. J. (2011). *Issues associated with vertical scales for PARCC assessments*. Retrieved from <http://www.parcconline.org/sites/parcc/files/PARCCVertScal289-12-201129.pdf>
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.
- La Salle, A., Munoz, C., Ruff, L., Weisman, E., Sedillo, R., & Phillips, L. (2012, April). *Grounded in the content: The role of content analysis in evidence-based standard setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.
- Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory* (CASMA Research Report No. 27). Iowa City, IA: University of Iowa.
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23(4), 6–15.
- Liang, T., Han, K. T., & Hambleton, R. K. (2008). *User's guide for ResidPlots-2: Computer software for IRT graphical residual analyses, Version 2.0* (Center for Educational Assessment Research Report No. 688). Amherst, MA: University of Massachusetts, Center for Educational Assessment.
- Liang, T., Han, K. T., & Hambleton, R. K. (2009). ResidPlots-2: Computer software for IRT graphical residual analyses. *Applied Psychological Measurement*, 33(5), 411–412.
- Linacre, J. M. (2004). *A user's guide to Winsteps Rasch-model computer programs*. Chicago, IL: MESA Press.
- 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.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 31(6), 3–16.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- 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, LA.
- Marion, S., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., Sheinker, J., & Council of Chief State School Officers (CCSSO). (2002). *Making valid and reliable decisions in determining adequate yearly progress. A paper in the series: Implementing the State Accountability System Requirements under the No Child Left Behind Act of 2001*. Washington, DC: CCSSO.

- Marsh, H. W., Martin, A. J., & Jackson, S. (2010). Introducing a short version of the physical self description questionnaire: New strategies, short-form evaluative criteria, and applications of factor analyses. *Journal of Sport & Exercise Psychology*, 32(4), 438–482.
- Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction. *Review of Educational Research*, 4, 1332–1361.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.
- Mislevy, R. J. (2009). *Validity from the perspective of model-based reasoning* (CRESST Report 752). Los Angeles, CA: National Center for Research on Evaluation, Standards, & 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.
- National Assessment Governing Board (NAGB). (2010). *Program of preparedness research study brief*. Retrieved from <http://www.nagb.org/content/nagb/assets/documents/newsroom/press-releases/2010/release-20101122/research.pdf>
- National Council on Measurement in Education (NCME). (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Madison, WI: Author.
- National Governors Association Center for Best Practices (NGA Center) & CCSSO. (2010). *Common Core State Standards for mathematics*. Washington, DC: Authors.
- Northwest Evaluation Association. (2005). *Technical manual: For use with Measures of Academic Progress and achievement level tests*. Lake Oswego, OR: Author.
- O'Malley, K., Keng, L., & Miles, J. (2012). Using validity evidence to set performance standards. In G. J. Cizek (Ed.), *Setting performance standards* (2nd ed.) (pp. 301–322). New York, NY: Routledge.
- O'Neil, T., Sireci, S. G., & Huff, K. F. (2004). Evaluating the consistency of test content across two successive administrations of a state-mandated science assessment. *Educational Assessment*, 9, 129–151.
- Patz, R. J. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: CCSSO.
- 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(4), 359–370.
- Pitoniak, M. J., Sireci, S. G., & Luecht, R. M. (2002). A multitrait-multimethod validity investigation of scores from a professional licensure exam. *Educational and Psychological Measurement*, 62, 498–516.
- 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 the American Educational Research Association, New Orleans, LA.
- Rabinowitz, S., Zimmerman, J., & Sherman, K. (2001). *Do high-stakes tests drive up student dropout rates? Myths versus realities* (Research brief). San Francisco, CA: WestEd.
- Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Erlbaum.
- Reckase, M. D. (2006a). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of the standard setting methods. *Educational Measurement: Issues and Practice*, 25(2), 4–18.
- Reckase, M. D. (2006b). Rejoinder: Evaluating standard setting methods using error models proposed by Schultz. *Educational Measurement: Issues and Practice*, 25(3), 14–17.
- Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, 3, 1–20.
- Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests*. Washington, DC: National Research Council.
- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7(14).
- Rudner, L. M. (2004). *Expected classification accuracy*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Select Committee. (1994/1995). Sampling and statistical procedures used in the California Learning Assessment System. In L. J. Cronbach (Ed.). *A valedictory: Reflections on 60 years in educational testing* (Board Bulletin). Washington, DC: National Research Council, Board on Testing and Assessment.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450.
- Silk, Y., Silver, D., Amerian, S., Nishimura, C., & Boscardin, C. K. (2009). *Using classroom artifacts to measure the efficacy of a professional development* (CRESST Report 761). Los Angeles, CA: National Center for Research on Evaluation, Standards, & Student Testing.
- Simpson, M., Gong, B., Marion, S., National Center on Educational Outcomes, CCSSO, & National Association of State Directors of Special Education. (2006). *Effect of minimum cell sizes and confidence interval sizes for special education subgroups on school-level AYP determinations* (Synthesis Report 61). Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.
- Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice*, 16(1), 12–19.
- Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299–321.

- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing Inc.
- Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement*, 16, 17–31.
- Sireci, S. G., & Geisinger, K. F. (1995). Using subject matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement*, 19, 241–255.
- Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment*, 13, 108–131.
- Sireci, S. G., Hauger, J. B., Wells, C. S., Shea, C., & Zenisky, A. L. (2009). Evaluation of the standard setting on the 2005 grade 12 National Assessment of Educational Progress mathematics test. *Applied Measurement in Education*, 22, 339–358.
- Sireci, S. G., & Mullane, L. A. (1994). Evaluating test fairness in licensure testing: The sensitivity review process. *CLEAR Exam Review*, 5(2), 22–28.
- Sireci, S. G., Robin, F., Meara, K., Rogers, H. J., & Swaminathan, H. (2000). An external evaluation of the 1996 grade 8 NAEP science framework. In N. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 74–100). Washington, DC: National Academies Press.
- Sireci, S. G., Rogers, H. J., Swaminathan, H., Meara, K., & Robin, F. (2000). Appraising the dimensionality of the 1996 grade 8 NAEP science assessment data. In N. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 101–122). Washington, DC: National Academies Press.
- Sireci, S. G., & Schweid, J. A. (2011, April). *Beyond alignment: Important questions to ask (and answer) to evaluate content validity*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Sireci, S. G., & Wells, C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33–68). Washington, DC: CCSSO.
- Smarter Balanced Assessment Consortium (Smarter Balanced). (2010). *Race to the Top assessment program application for new grants: Comprehensive assessment systems, CFDA Number: 84.395B*. OMB Control Number 1810-0699.
- Smarter Balanced Assessment Consortium (Smarter Balanced). (2012a). *Master work plan—formative*. Retrieved from <http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/03/Formative-Assessment-Master-Work-Plan-Narrative.pdf>
- Smarter Balanced Assessment Consortium (Smarter Balanced). (2012b). *Theory of action. An excerpt from the Smarter Balanced Race to the Top application*. Retrieved from <http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/02/Smarter-Balanced-Theory-of-Action.pdf>

- Smarter Balanced Assessment Consortium (Smarter Balanced). (n.d.). *Frequently asked questions*. Retrieved from <http://www.smarterbalanced.org/resources-events/faqs/>
- Tomlinson, M. R., & Fortenberry, N. (2008, October). *Classroom artifacts: Tools to assess the use of active, innovative, and engineering pedagogies among engineering faculty*. Paper presented at the annual ASEE/IEEE Frontiers in Education conference, Saratoga Springs, NY. Retrieved from <http://fie-conference.org/fie2008/papers/1088.pdf>
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement In Education*, 20(2), 227–253.
- U.S. Department of Education. (2009a). *Race to the Top program executive summary*. Washington, DC: Author.
- U.S. Department of Education. (2009b). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001. [Revised December 21, 2007 to include modified academic achievement standards. Revised with technical edits, January 12, 2009]*. Washington, DC: Author.
- U.S. Department of Education (2010). *U.S. secretary of education Duncan announces winners of competition to improve student assessments*. Retrieved from <http://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse>
- Vasavada, N., Carman, E., Hart, B., & Luisser, D. (2010). *Common Core State Standards alignment: ReadStep, PSAT/NMSQT, and SAT* (Research Report 2010-5). New York, NY: The College Board.
- Wainer, H. (2011). *Uneducated guesses: Using evidence to uncover misguided education policies*. Princeton, NJ: Princeton University Press.
- Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36, 301–335.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Washington, DC: CCSSO.
- Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7–25.
- Welch, C., & Dunbar, S. (2011, April). *K–12 assessments and college readiness: Necessary validity evidence for educators, teachers, and parents*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Wells, C. S., Baldwin, S., Hambleton, R. K., Sireci, S. G., Karantonis, A., & Jirka, S. (2009). Evaluating score equity assessment for state NAEP. *Applied Measurement in Education*, 22, 394–408.
- Wilson, D, Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis* [computer program]. Mooresville, IN: Scientific Software.

- Williams, N. J., Keng, L., & O'Malley, K. (2012, April). *Maximizing panel input: Incorporating empirical evidence in a way the standard setting panel will understand*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.
- Wyatt, J., Kobrin, J., Wiley, A., Camara, W. J., & Proestler, N. (2011). *SAT benchmarks: Development of a college readiness benchmark and its relationship to secondary and postsecondary school performance*. (Research Report 2011-5). New York, NY: The College Board.
- Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.
- Yen, W. M. (1997). The technical quality of performance assessments: Standard errors of percents of pupils reaching standards. *Educational Measurement: Issues and Practice*, 16(3), 5–15.
- Young, J., Pitoniak, M. J., King, T. C., & Ayad, E. (2012). *Smarter Balanced Assessment Consortium: Guidelines for accessibility for English language learners*. Retrieved from <http://www.smarterbalanced.org/smarter-balanced-assessments/>
- Zwick, R., & Schlemer, L. (2004). SAT validity for linguistic minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice*, 23(1), 6–16.

Appendix A: Smarter Balanced Theory of Action and Derivation of Purpose Statements

Smarter Balanced Assessment Consortium Theory of Action

Bennett (2010) described a Theory of Action (TOA) as follows:

Theory of Action is a common notion in the program evaluation literature . . . appearing to have come about because program managers were too often unclear about the intended goals of their efforts. The term is closely associated with *logic model*, a graphical or textual description of an intervention that explains the cause-effect relationships among inputs, activities, and intended outcomes. (pp. 70-71)

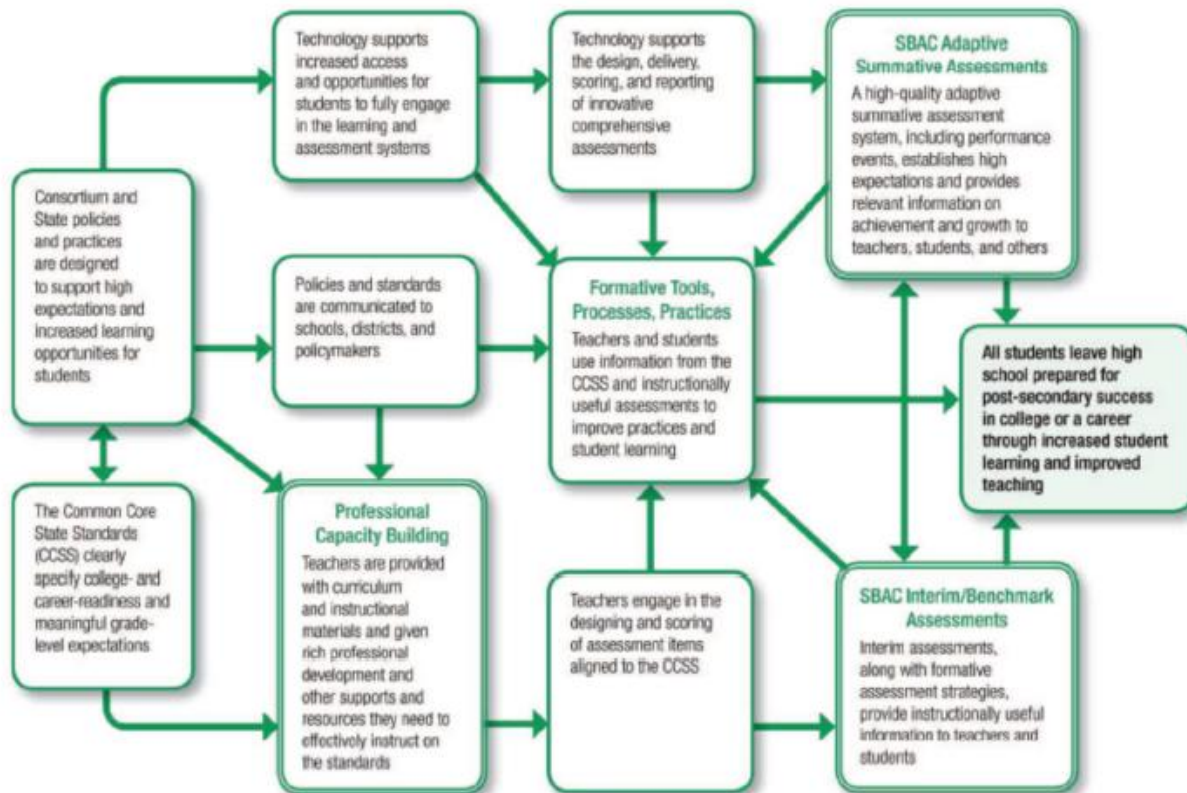
Smarter Balanced's TOA is well articulated in its Race to the Top application (Smarter Balanced, 2010) and has been excerpted from the application as a separate document available on the SBAC website (Smarter Balanced, 2012b). It begins by stating that Smarter Balanced "supports the development and implementation of learning and assessment systems to radically reshape the education enterprise . . . to improve student outcomes" and states that "the overarching goal of the Smarter Balanced Assessment Consortium is to *ensure that all students leave high school prepared for postsecondary success in college or a career through increased student learning and improved teaching*" (p. 1; emphasis in original). The TOA lists "seven principles undergirding the theory of action" (p. 1). These principles are:

1. Assessments are grounded in a thoughtful, standards-based curriculum and are managed as part of an integrated system of standards, curriculum, assessment, instruction, and teacher development.
2. Assessments produce evidence of student performance on challenging tasks that evaluate the Common Core State Standards.
3. Teachers are integrally involved in the development and scoring of assessments.
4. The development and implementation of the assessment system is a state-led effort with a transparent and inclusive governance structure.
5. Assessments are structured to continuously improve teaching and learning.
6. Assessment, reporting, and accountability systems provide useful information on multiple measures that is educative for all stakeholders.
7. Design and implementation strategies adhere to established professional standards. (Smarter Balanced, 2010, pp. 32–33)

From these principles we can immediately infer that intended goals of Smarter Balanced are to develop quality assessments that are aligned with the CCSS, are part of a system that supports instruction and student learning, and provide results that are useful for evaluating student performance. It is also clear that other goals are to involve teachers throughout the test development and scoring processes and to operate as a true collaborative with states working in unison toward these common goals.

The model that Smarter Balanced established to meet these goals involves three different components: (a) summative assessments, (b) interim-benchmark assessments, and (c) formative assessment resources. A schematic representation of the Smarter Balanced TOA is illustrated in Figure A-1, which is taken directly from the Smarter Balanced Race to the Top application (Smarter Balanced, 2010). This representation includes the three assessment components, but also illustrates the other components that are required for the Consortium members to work together in unison and to reach the "overarching goal" found on the right side of the figure. Related to the Theory of Action are the overall and specific claims for the summative assessments, which are presented in Table A-1.

Figure A-1. Overview of Smarter Balanced Assessment Consortium Theory of Action



Source: Smarter Balanced (2012b).

Table A-1. Overall and Specific Claims for Smarter Balanced Summative Assessments

Claim Type	ELA: Students can . . .	Mathematics: Students can . . .
Overall: Grades 3–8	demonstrate progress toward college and career readiness in English language arts and literacy.	demonstrate progress toward college and career readiness in mathematics.
Overall: Grade 11	demonstrate college and career readiness in English language arts and literacy.	demonstrate college and career readiness in mathematics.
Specific	read closely and analytically to comprehend a range of increasingly complex literary and informational texts.	explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
	produce effective and well-grounded writing for a range of purposes and audiences.	solve a range of complex, well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies.
	employ effective speaking and listening skills for a range of purposes and audiences.	clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.
	engage in research and inquiry to investigate topics, and to analyze, integrate, and present information.	analyze complex, real-world scenarios and construct and use mathematical models to interpret and solve problems.

Appendix B: Description of Alignment Methods

Alignment Model	Dimension	Brief Description
Webb (1997)	*Categorical Concurrence	Match of items to general content areas
	**Depth of Knowledge Consistency	Cognitive level of items compared to cognitive level of benchmark/objective
	**Range of Knowledge Correspondence	Number of benchmarks/objectives measured within general content area
	**Balance of Representation	Distribution of items across general content areas
Achieve (2006)	*Content Centrality	Congruence between item and objective/benchmark
	*Performance Centrality	Congruence between cognitive demand of item and objective/benchmark
	**Source of Challenge	Grade-level appropriateness
	*Level of Cognitive Demand	Cognitive level measured by item
	**Level of Challenge	Degree to which test captures difficulty implied by general content areas
	**Balance	Holistic evaluation of how well test represents content/cognitive specs
	**Range	Proportion of objectives/benchmarks measured within general content area
SEC (Porter et al., 2001)	*Content Match	Match of items to content areas and cognitive levels
	**Expectations for Student Performance	Compares cognitive demands of curriculum and assessment
	**Instructional Content	Compares what is taught with what is tested

*Covered or partially covered by one or more traditional content validation approaches.

**Unique contribution of alignment method.

From Sireci & Schweid (2011).

Appendix C: Description of Item Similarity Rating Approach to Evaluating Test Content

As stated earlier, a disadvantage of this approach to blueprint confirmation is that it may foster social desirability—that is, by informing SMEs of the intended CCSS measured by each item, it may unconsciously bias their ratings in support of item/standard congruence. To avoid this potential confound, and to determine whether other relations among the items are present that are not described in the test specifications, the item similarity rating task described earlier could be conducted. An example of this task is presented in Figure C-1. An example of some of the results from this type of study (from Sireci, Robin, Meara, Rogers, & Swaminathan, 2000) is presented in Figure C-2. These results could be followed up by cluster analyses, to see if the items cluster as intended by the test specifications.

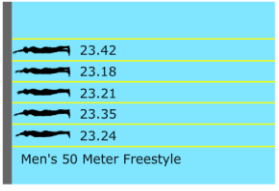
Given that the item similarity rating task requires more SME time and more complex data analysis, we recommend that all items be rated for congruence using an alignment-type rating task similar to that illustrated in Exhibit 1. However, the similarity rating procedure provides a more stringent test and protects against confirmationist bias (social desirability), and so should be considered as a supplementary study, perhaps using a subset of items.

Figure C-1. Example of Item Similarity Rating Task

Directions: Please review each pair of items and rate how similar the two items are to one another in terms of the mathematics knowledge and skills measured using the rating scale provided.

43025

Five swimmers compete in the 50-meter race. The finish time for each swimmer is shown in the video.



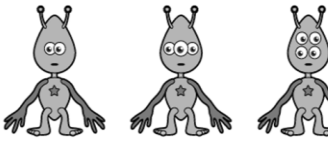
Men's 50 Meter Freestyle

23.42
23.18
23.21
23.35
23.24

Explain how the results of the race would change if the race used a clock that rounded to the nearest tenth.

43081

The two-eyed space creatures, three-eyed space creatures, and four-eyed space creatures are having a contest to create a group with 24 total eyes.



How many two-eyed space creatures are needed to make a group with 24 total eyes?

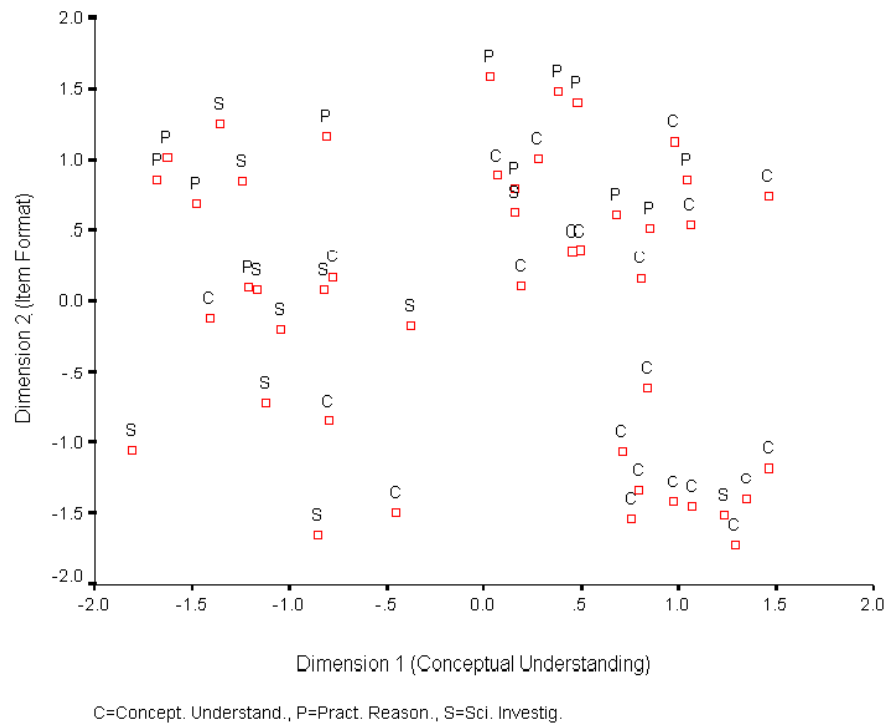
1 2 3
4 5 6
7 8 9
0

Delete

1
2
3
4
5
6
7
8

Very Similar
Very Different

Figure C-2. Example of Results from Item Similarity Ratings Study



Source: Sireci et al., 2000.

Appendix D: Description of ResidPlots2: IRT Residual Analysis Software

ResidPlots-2 (Liang, Han, & Hambleton, 2008, 2009) is a software program for evaluating the fit of item response theory (IRT) models to data. By comparing observations to model-predicted expectations, ResidPlots-2 works at the item level to provide researchers with information to determine how well an IRT model fits a given data set. The approach used in ResidPlots-2 is to first compute model fit statistics using the observed data, and then also use item and ability estimates from IRT estimation programs, such as BILOG-MG, PARSCALE, and MULTILOG, to simulate examinee response data and report the average from 10 replications of the simulation. Thus, simulation results obtained in this way better approximate the expected observed test score distribution.

The output from ResidPlots-2 takes the forms of both graphs and tables. Plots generated by ResidPlots-2 include:

- Item-level plots (raw residual plots, standardized residual plots),
- Test-level plots (standardized residual distributions [both cumulative density function {CDF} and probability density function {PDF}], item and score fit plots from empirical and simulated data); and
- Score plots (observed and predicted test score distributions).

ResidPlots-2 also generates six tables of results:

- The *FIT STAT* table provides results for two fit statistics at the item level (chi square, G square) as well as degree of freedom and fit probability for both, and basic item details (item number, parameter estimates, and sample size).
- The *SR PDF* table lists details of the standardized residual (SR) distribution for the PDF, with mean, standard deviation, and relative frequency of the SR distribution. These results are provided for the overall test and broken out by format (dichotomous and polytomous items) and for both observed and simulated data.
- The *SR CDF* table is a companion table to the SR PDF table; here, the results are provided for the CDF.
- The *NCOUNT* table displays the characteristics of the sample (sample size and percentage) in each reported interval for each item. This is an important feature, as users can make application-specific choices about interval width and score ranges in ResidPlots-2.
- The *PFIT* table provides the results of the Lz person fit statistic for each person in the sample. Note that this report lists the probability values for each person, where values below 0.05 are indicative of person misfit.
- The *P_RISE* table contains results for the root integrated square error statistic (RISE), which is a nonparametric fit statistic. As with the PFIT table, results are shown in terms of probability values for each item, where values less than 0.05 are indicative of nonparametric item misfit.

The plots in Figures D-1 and D-2 are samples of output from ResidPlots-2 that depict the item-fit plot. Note that the 3P model was fit to the data for Figure D-1, while a 1P model was fit to the same data for Figure D-2. Figure D-2 illustrates that results from the observed calibration are much more disparate from the simulated results than the results shown in Figure D-1, which suggests that the 3P model provides better model-data fit than the 1P model for the data.

Figure D-1. ResidPlots-2 Item Fit Plot (data fit by 3P)

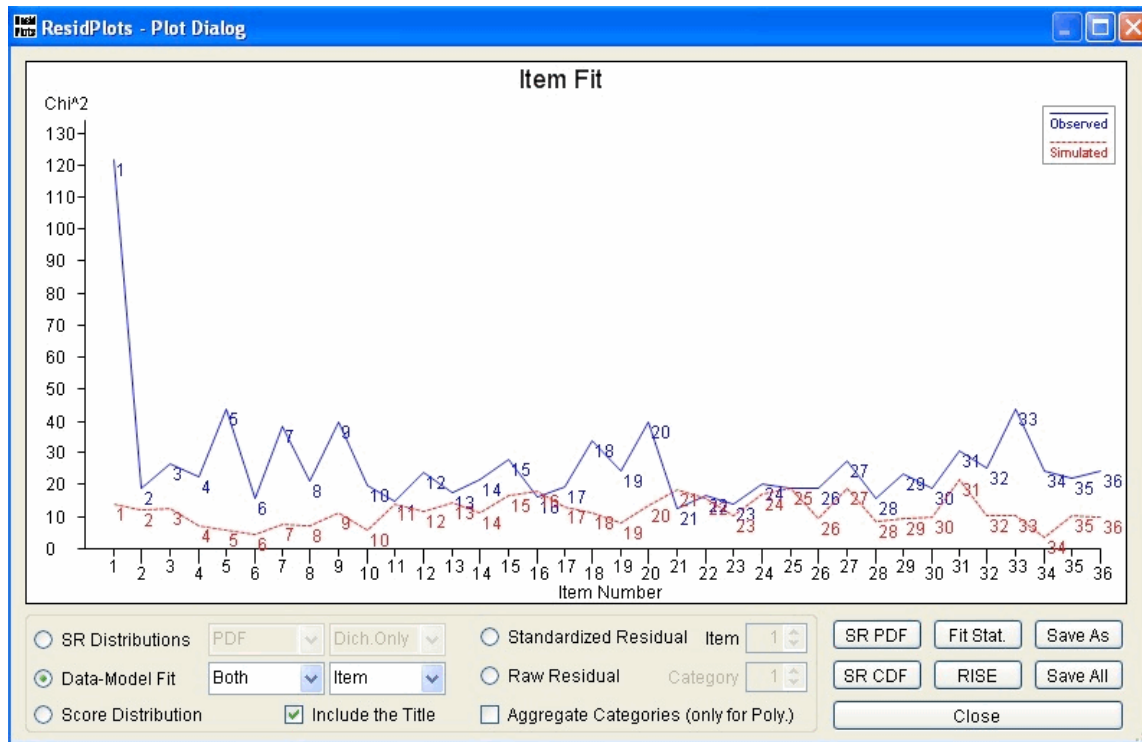


Figure D-2. ResidPlots-2 Item Fit Plot (data fit by 1P)

