

Chapter 5 Test Fairness 2

 Introduction..... 2

 Definitions for Validity, Bias, Sensitivity, and Fairness. 2

 The Smarter Balanced Accessibility and Accommodations Framework 4

 Figure 1. Conceptual Framework for Usability, Accessibility, and Accommodations.7

 Meeting the Needs of Traditionally Underrepresented Populations. 7

 The Individual Student Assessment Accessibility Profile (ISAAP). 8

 Usability, Accessibility, and Accommodations Guidelines: Intended Audience and Recommended Applications..... 10

 Guidelines for Accessibility for English Language Learners. 11

 Fairness as a Lack of Measurement Bias: Differential Item Functioning Analyses 13

 Test Fairness and Implications for Ongoing Research..... 13

 Internal Structure. 14

 Response Processes. 14

 Relationships with Other Variables. 14

 Test Consequences. 15

 References..... 16

Chapter 5 Test Fairness

Introduction

For large-scale programs such as those developed by the Smarter Balanced Assessment Consortium (Smarter Balanced), an essential goal is to ensure that all students have comparable opportunities to demonstrate their achievement level. Smarter Balanced strives to provide every student with a positive and productive assessment experience and results that are a fair and accurate depiction of each student's achievement. Ensuring test fairness is a fundamental part of validity, starting with test design, and is an important feature built into each step of the test development process, such as item writing, test administration, and scoring. The 2014 *Standards for Educational and Psychological Testing* state, "The term fairness has no single technical meaning, and is used in many ways in public discourse." It also suggests that fairness to all individuals in the intended population is an overriding and fundamental validity concern.

The Smarter Balanced system is designed to provide a valid, reliable, and fair measure of student achievement based on the Common Core State Standards (CCSS). The validity and fairness of the measures of student achievement are influenced by a multitude of factors; central among them are:

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

Smarter Balanced uses several documents to address reliability, validity, and fairness. The *Common Core State Standards were originally* developed by the National Governors Association (NGA) and the Council Chief State School Officers (CCSSO). The *Smarter Balanced Content Specifications*, developed by the Consortium, articulate the claims and targets of the Smarter Balanced assessments, defining the knowledge, skills, and abilities to be assessed and their relationship to the CCSS. In doing so, these documents describe the major constructs—identified as "Claims"—within English language arts/literacy (ELA/literacy) and mathematics for which evidence of student achievement will be gathered and which will form the basis for reporting student performance. Much of the evidence presented in this chapter pertains to fairness in treatment during the testing process and lack of measurement bias (i.e., DIF). Fairness (minimizing bias) and the design of accessibility supports (i.e., universal tools, designated supports and accommodations in content development is addressed in Chapter 3.

Definitions for Validity, Bias, Sensitivity, and Fairness. Some key concepts for the ensuing discussion concern validity, bias, and fairness and are described as follows.

Validity. Validity is the extent to which the inferences and actions made based on test scores are appropriate and backed by evidence (Messick, 1989). It constitutes the central notion underlying the development, administration and scoring of a test, as well as the uses and interpretations of test scores. Validation is the process of accumulating evidence to support each proposed score interpretation or use. Evidence in support of validity is extensively discussed in Chapter 2.

Bias and sensitivity. According to the Standards for Educational and Psychological Testing, “Bias in tests and testing refers to construct-irrelevant [i.e., invalid] components that result in systematically lower or higher scores for identifiable groups of examinees” (*Standards*, 1999 (AERA, APA, & NCME, p. 76; *Standards*, (AERA, APA, & NCME, 2014, 51-54). “Sensitivity” is used to refer to an awareness of the need to avoid bias in assessment. In common usage, reviews of tests for bias and sensitivity are reviews to help ensure that the test items and stimuli are fair for various groups of test takers, (*Standards*, 2014 (AERA, APA, & NCME, 2014, p. 64).

The goal of fairness in assessment can be approached by ensuring that test materials are as free as possible of unnecessary barriers to the success of a diverse group of students. Smarter Balanced developed *Bias and Sensitivity Guidelines* to help ensure that the assessments are fair for all groups of test takers, despite differences in characteristics including, but not limited to, disability status, ethnic group, gender, regional background, native language, race, religion, sexual orientation, and socioeconomic status. Unnecessary barriers can be reduced by following some fundamental rules (ETS, 2012):

- not measuring irrelevant knowledge or skills (i.e., construct irrelevant),
- not angering, offending, upsetting, or otherwise distracting test takers, and
- treating all groups of people with appropriate respect in test materials.

These rules help ensure that the test content is fair for test takers as well as acceptable to the many stakeholders and constituent groups within the Smarter Balanced states. The more typical view is that bias and sensitivity guidelines apply primarily to the review of test items. However, fairness must be considered in all phases of test development and use. Smarter Balanced strongly relied on the *Bias and Sensitivity Guidelines* in the development of the Smarter Balanced assessments, particularly in item writing and review. Items had to comply with the *bias and sensitivity Guidelines* in order to be included in the Smarter Balanced assessments. Use of the *Guidelines* will help the Smarter Balanced assessments comply with Chapter 3, Standard 3.2 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p. 82). Standard 3.2 states that “Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests’ being affected by construct-irrelevant characteristics such as linguistic, communicative, cognitive, cultural, physical or other characteristics”. The Smarter Balanced assessments were developed using the principles of evidence-centered design (ECD). Three basic elements of ECD (Mislevy, Steinberg, & Almond, 1999) are stating the claims to be made about test takers, deciding what evidence is required to support these claims, and administering test items that provide the required evidence. ECD provides a chain of evidence-based reasoning that links test performance to the Claims to be made about test takers. Fair assessments are essential to the implementation of ECD. If the items are not fair, then the evidence they provide means different things for the various groups of students. Under those circumstances, the Claims cannot be equally supported for all test takers, which is a threat to validity. Appropriate use of the *Bias and Sensitivity Guidelines* helps to

ensure that the evidence provided by the items allows ECD to function as intended and is equally valid for various groups of test takers.

Fairness. “Fairness” as mentioned previously is a difficult word to define because, as indicated in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, p. 74), “The central idea of fairness in testing is to identify and remove construct-irrelevant barriers to maximal performance for any examinee.” An extensive discussion of the meanings of “fairness” in assessment was also given by Camilli (2006). A useful definition of fairness for the purposes of the *Bias and Sensitivity Guidelines* is the extent to which the test scores are valid for different groups of test takers. For example, a mathematics item may contain difficult language unrelated to mathematics. If the language interfered about equally with all test takers, validity would be negatively impacted for all test takers. If, however, the language were a more significant barrier for students who are not native speakers of English, compared with other students, then the item would be unfair. If items are more difficult for some groups of students than for other groups of students, the items may not necessarily be unfair. For example, if an item were intended to measure the ability to comprehend a reading passage in English, score differences between groups based on real differences in comprehension of English would be valid and, therefore, fair. As Cole and Zieky (2001, p. 375) noted, “If the members of the measurement community currently agree on any aspect of fairness, it is that score differences alone are not proof of bias.” Fairness does not require that all groups have the same average scores. Fairness requires any existing differences in scores to be valid. An item would be unfair if the source of the difficulty were not a valid aspect of the item. For example, an item would be unfair if members of a group of test takers were distracted by an aspect of the item that they found highly offensive. If the difference in difficulty reflected real and relevant differences in the group’s level of mastery of the tested CCSS, the item could be considered as fair.

The Smarter Balanced Accessibility and Accommodations Framework

Smarter Balanced has built a framework of accessibility for all students, including English Language Learners (ELLs), students with disabilities, and ELLs with disabilities, but not limited to those groups. Three additional sources—the Smarter Balanced *Item Specifications*, the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* and the Smarter Balanced *Bias and Sensitivity Guidelines*—are used to guide the development of items and tasks to ensure that they accurately measure the targeted constructs. Recognizing the diverse characteristics and needs of students who participate in the Smarter Balanced assessments, the states worked together through the Smarter Balanced Test Administration and Student Access Work Group to develop an Accessibility and Accommodations Framework that guided the Consortium as it worked to reach agreement on the specific universal tools, designated supports, and accommodations available for the assessments. This work also incorporated research and practical lessons learned through Universal Design, accessibility tools, and accommodations (Thompson, Johnstone, & Thurlow, 2002). Much of the conceptualization for this chapter is a direct reflection of the outcomes from the work of the Test Administration and Student Access Work Group.

In the process of developing its next-generation assessments to measure students’ knowledge and skills as they progress toward college and career readiness, Smarter Balanced recognized that the validity of assessment results depends on each student having appropriate universal tools, designated supports, and/or accommodations when needed, based on the constructs being measured by the assessment. The

Smarter Balanced Assessment System utilizes technology intended to deliver assessments that meet the needs of individual students. Online/electronic delivery of the assessments helps ensure that students are administered a test individualized to meet their needs consistent with their peers. Items and tasks were delivered using a variety of accessibility resources and accommodations that can be administered to students automatically based on their individual profiles. Accessibility resources include but are not limited to foreground and background color flexibility, tactile presentation of content (e.g., Braille), and translated presentation of assessment content in signed form and selected spoken languages.

A principle for Smarter Balanced was to adopt a common set of accessibility resources and accommodations. Moreover, the Notification Inviting Applications (NIA) posted in the Federal Register, April 9, 2010, required “a common set of policies and procedures for accommodations” for any consortia funded by the USED Race to the Top Assessment Program; the following definition was used.

Accommodations means changes in the administration of an assessment, including but not limited to changes in assessment setting, scheduling, timing, presentation format, response mode, and combinations of these changes that do not change the construct intended to be measured by the assessment or the meaning of the resulting scores. Accommodations must be used for equity in assessment and not provide advantage to students eligible to receive them.

The focus is on “equity in assessment” and does not refer to specific student characteristics, a perspective that is consistent with the Accessibility and Accommodations Framework. A fundamental goal was to design an assessment that is accessible for all students, regardless of English language proficiency, disability, or other individual circumstances. The three components of the *Accessibility and Accommodations Framework* are designed to meet that need. The intent was to ensure that the following steps were achieved for Smarter Balanced.

- Design and develop items and tasks to ensure that all students have access to the items and tasks designed to measure the targeted constructs. In addition, deliver items, tasks, and the collection of student responses in a way that maximizes validity for each student.
- Adopt the conceptual model embodied in the Accessibility and Accommodations Framework that describes accessibility resources of digitally delivered items/tasks and acknowledges the need for some adult-monitored accommodations. The model also characterizes accessibility resource as a continuum from those available to all students ranging to ones that are implemented under adult supervision available only to those students with a documented need.
- Implement the use of an individualized and systematic needs profile, or Individual Student Assessment Accessibility Profile (ISAAP), for students that promotes the provision of appropriate access and tools for each student. Smarter created an ISAAP process that helps education teams systematically select the most appropriate accessibility resources for each student and ISAAP tool, which helps teams note the accessibility resources chosen.

The conceptual framework that serves as the basis underlying the usability, accessibility, and accommodations is shown in Figure 1. This figure portrays several aspects of the Smarter Balanced assessment features—universal tools (available for all students), designated supports (available when

indicated by an adult or team), and accommodations as documented in an Individualized Education Program (IEP) or 504 plan. It also displays the additive and sequentially inclusive nature of these three aspects. Universal tools are available to all students, including those receiving designated supports and those receiving accommodations. Designated supports are available only to students who have been identified as needing these accommodations (as well as those students for whom the need is documented). Accommodations are available only to those students with documentation of the need through a formal plan (e.g., IEP). Those students also may access designated supports and universal tools.

A universal tool for a content focus in a specific may be an accommodation for another grade or content focus. Similarly, a designated support may also be an accommodation, depending on the content target and grade. This approach is consistent with the emphasis that Smarter Balanced has placed on the validity of assessment results coupled with access. Universal tools, designated supports, and accommodations are all intended to yield valid scores. Universal tools, designated supports, and accommodations result in scores that count toward participation in statewide assessments. Also shown in Figure 1 are the universal tools, designated supports, and accommodations for each category of accessibility resources. There are both embedded and non-embedded versions of the universal tools, designated supports, or accommodations depending on whether they are provided as digitally delivered components of the test administration or separate from test delivery.

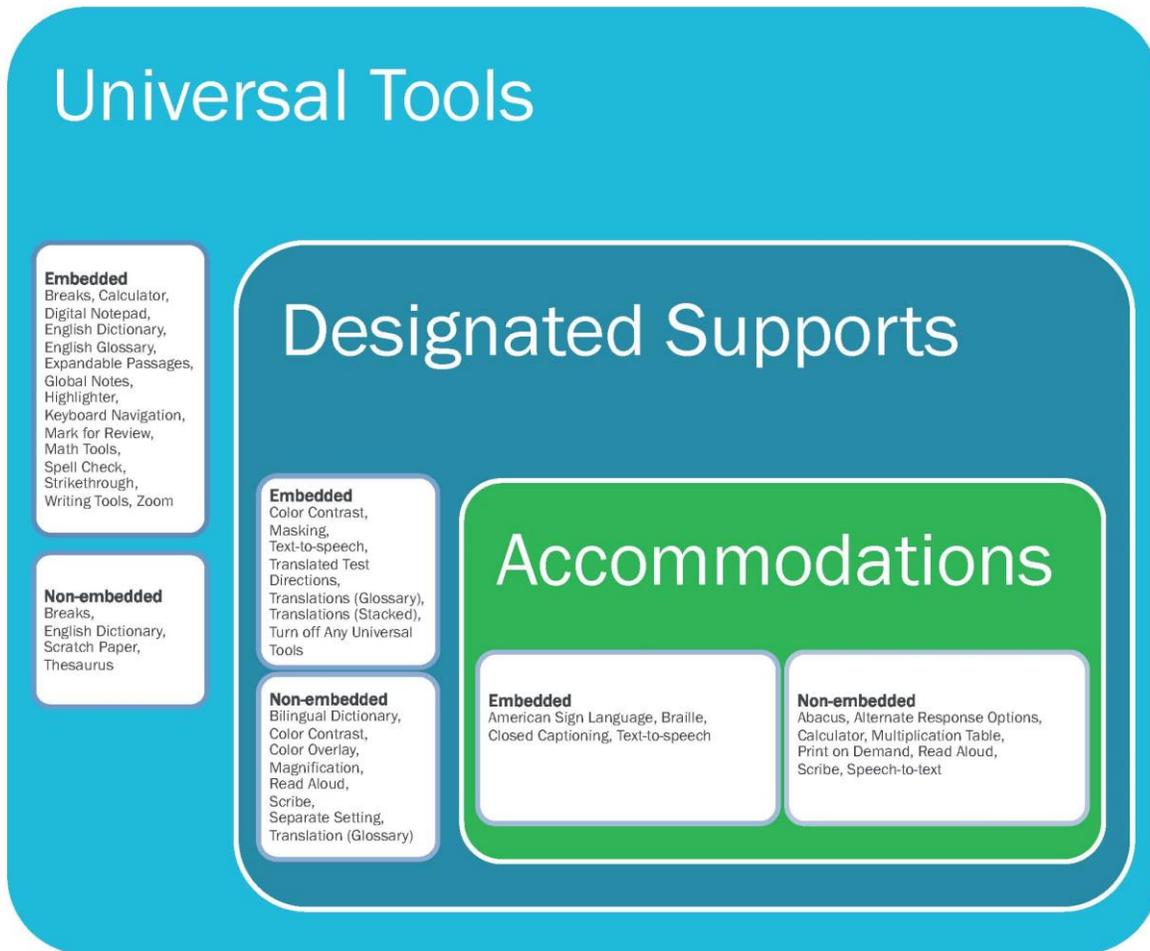


Figure 1. Conceptual Framework for Usability, Accessibility, and Accommodations.

Meeting the Needs of Traditionally Underrepresented Populations. The policy decision was to make accessibility resources available to all students based on need rather than eligibility status or student subgroup categorical designation. This reflects a belief among Consortium states that unnecessarily restricting access to accessibility resources threatens the validity of the assessment results and places students under undue stress and frustration. Additionally, accommodations are available for students who qualify for them. Although the intention of this policy is to ensure a positive and productive assessment experience for all students, elimination of specific eligibility criteria may raise concerns among some educators and advocates who worked to create eligibility criteria that guarantee appropriate assessment supports for their students of interest. Discussion on how a needs-based approach will benefit ELLs, students with disabilities, and ELLs with disabilities is presented here.

How the Framework Meets Needs of Students Who Are ELLs. Students who are ELLs have needs that are unique from those students with disabilities, including language-related disabilities. The needs of ELLs are not the result of a language-related disability, but instead are specific to the student's current level of English language proficiency. The needs of students who are ELLs are diverse and are influenced by the interaction of several factors, including their current level of English language proficiency, their prior exposure to academic content and language in their native language, the languages to which they are exposed outside of school, the length of time they have participated in the U.S. education system, and the language(s) in which academic content is presented in the classroom. Given the unique background and needs of each student, the conceptual framework is designed to focus on students as individuals and to provide several accessibility resources that can be combined in a variety of ways. Some of these digital tools, such as using a highlighter to highlight key information and an audio presentation of test navigation features, are available to all students, including those at various stages of English language development. Other tools, such as the audio presentation of items and glossary definitions in English, may also be assigned to any student, including those at various stages of English language development. Still other tools, such as embedded glossaries that present translation of construct irrelevant terms, are intended for those students whose prior language experiences would allow them to benefit from translations into another spoken language. Collectively, the conceptual framework for usability, accessibility, and accommodations embraces a variety of accessibility resources that have been designed to meet the needs of students at various stages in their English language development.

How the Framework Meets Needs of Students with Disabilities. Federal law requires that students with disabilities who have a documented need receive accommodations that address those needs, and that they participate in assessments. The intent of the law is to ensure that all students have appropriate access to instructional materials and are held to the same high standards. When students are assessed, the law ensures that students receive appropriate accommodations during testing so they can appropriately demonstrate what they know and can do so that their achievement is measured accurately.

The Accessibility and Accommodations Framework addresses the needs of students with disabilities in three ways. First, it provides for the use of digital test items that are purposefully designed to contain multiple forms of the item, each developed to address a specific access need. By allowing the delivery of a given access form of an item to be tailored based on each student's access need, the Framework fulfills the intent of Federal accommodation legislation. Embedding universal accessibility digital tools, however, addresses only a portion of the access needs required by many students with disabilities. Second, by embedding accessibility resources in the digital test delivery system, additional access needs are met. This approach fulfills the intent of the law for many, but not all, students with disabilities, by allowing the accessibility resources to be activated for students based on their needs. Third, by allowing for a wide variety of digital and locally provided accommodations (including physical arrangements), the Framework addresses a spectrum of accessibility resources appropriate for math and ELA assessment. Collectively, the Framework adheres to Federal regulations by allowing a combination of universal design principles, universal tools, designated supports and accommodations to be embedded in a digital delivery system and through local administration assigned and provided based on individual student needs.

The Individual Student Assessment Accessibility Profile (ISAAP). Typical practice frequently required schools and educators to document, a priori, the need for specific student accommodations and then to document

the use of those accommodations after the assessment. For example, most programs require schools to document a student's need for a large-print version of a test for delivery to the school. Following the test administration, the school documented (often by bubbling in information on an answer sheet) which of the accommodations, if any, a given student received, whether the student actually used the large-print form, and whether any other accommodations, such as extended time, were provided. Traditionally, many programs have focused only on those students who have received accommodations and thus may consider an accommodation report as documenting accessibility needs. The documentation of need and use establishes a student's accessibility needs for assessment.

For most students, universal digital tools will be available by default in the Smarter Balanced test delivery system and need not be documented. These tools can be deactivated if they create an unnecessary distraction for the student. Other embedded accessibility resources that are available for any student needing them must be documented prior to assessment. Smarter Balanced intends to obtain information on individual student test administration conditions for students with specific accessibility needs not addressed in the *Usability, Accessibility, and Accommodations Guidelines*. To capture specific student accessibility needs, the Smarter Balanced Assessment System has established an individual student assessment accessibility profile (ISAAP). The ISAAP tool is designed to facilitate selection of the universal tools, designated supports and accommodations that match student access needs for the Smarter Balanced assessments, as supported by the *Smarter Balanced Usability, Accessibility, and Accommodations Guidelines*. The ISAAP Tool should be used in conjunction with the *Smarter Balanced Usability, Accessibility and Accommodations Guidelines* and state regulations and policies related to assessment accessibility as a part of the ISAAP process. For students requiring one or more accessibility resource, schools will be able to document this need prior to test administration. Furthermore, the ISAAP can include information about universal tools that may need to be eliminated for a given student. By documenting need prior to test administration, a digital delivery system will be able to activate the specified options when the student logs in to an assessment. In this way, the profile permits educators and schools to focus on each individual student, documenting the accessibility resources required for valid assessment of that student in a way that is efficient to manage.

The conceptual framework (Figure 1) provides a structure that assists in identifying which accessibility resources should be made available for each students. In addition, the conceptual framework is designed to differentiate between universal tools available to all students and accessibility resources that must be assigned before the administration of the assessment. Consistent with recommendations from Shafer and Rivera (2011), Thurlow, Quenemoen, and Lazarus (2011), Fedorchak (2012), and Russell (2011b), Smarter Balanced is encouraging schools to use a team approach to make decisions concerning each student's ISAAP. Gaining input from individuals with multiple perspectives, including the student, will likely result in appropriate decisions about the assignment of accessibility resources. Consistent with these recommendations avoidance of selecting too many accessibility resources for a student. The use of too many unneeded accessibility resources can decrease student performance.

The team approach encouraged by Smarter Balanced does not require the formation of a new decision-making team, and the structure of teams can vary widely depending on the background and needs of a student. A locally convened student support team can potentially create the ISAAP. For most students who do not require accessibility tools or accommodations, an initial decision by a teacher may be confirmed by a

second person (potentially the student). In contrast, for a student who is an English language learner and has been identified with one or more disabilities, the IEP team should include the English language development specialist who works with the student, along with other required IEP team members and the student, as appropriate. The composition of teams is not being defined by Smarter Balanced; it is under the control of each school and is subject to state and Federal requirements.

Usability, Accessibility, and Accommodations Guidelines: Intended Audience and Recommended Applications. The Smarter Balanced Consortium has developed *Usability, Accessibility, and Accommodations Guidelines* (UUAG) that are intended for school-level personnel and decision-making teams, particularly Individualized Education Program (IEP) teams, as they prepare for and implement the Smarter Balanced assessment. The UUAG provide information for classroom teachers, English development educators, special education teachers, and related services personnel to use in selecting and administering universal tools, designated supports, and accommodations for those students who need them. The UUAG are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment. The Smarter Balanced *Usability*, UUAG emphasize an individualized approach to the implementation of assessment practices for those students who have diverse needs and participate in large-scale content assessments. This document focuses on universal tools, designated supports, and accommodations for the Smarter Balanced content assessments of ELA/literacy and mathematics. At the same time, it supports important instructional decisions about accessibility for students who participate in the Smarter Balanced assessments. It recognizes the critical connection between accessibility in instruction and accessibility during assessment. The UUAG are also incorporated into the Smarter Balanced Test Administration Manual.

All students (including students with disabilities, ELLs, and ELLs with disabilities) are to be held to the same expectations for participation and performance on state assessments. Specifically, all students enrolled in grades 3 to 8 and 11 are required to participate in the Smarter Balanced mathematics except students with the most significant cognitive disabilities who meet the criteria for the mathematics alternate assessment based on alternate achievement standards (approximately 1% or less of the student population).

All students enrolled in grades 3 to 8 and 11 are required to participate in the Smarter Balanced English language/literacy assessment except:

- students with the most significant cognitive disabilities who meet the criteria for the English language/literacy alternate assessment based on alternate achievement standards (approximately 1% or fewer of the student population), and
- ELLs who are enrolled for the first year in a U.S. school. These students will participate in their state's English language proficiency assessment.

Federal laws governing student participation in statewide assessments include the Elementary and Secondary Education Act (ESEA)—reauthorized as the No Child Left Behind Act (NCLB) of 2001, the Individuals with Disabilities Education Improvement Act of 2004 (IDEA), and Section 504 of the Rehabilitation Act of 1973 (reauthorized in 2008).

Since the Smarter Balanced assessment is based on the CCSS, the universal tools, designated supports, and accommodations that are appropriate for the Smarter Balanced assessment may be different from those that state programs utilized previously. For the summative assessments, state participants can only make available to students the universal tools, designated supports, and accommodations consistent with the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines*. When the implementation or use of the universal tool, designated support, or accommodation is in conflict with a member state's law, regulation, or policy, a state may elect not to make it available to students.

The Smarter Balanced universal tools, designated supports, and accommodations currently available for the Smarter Balanced assessments have been prescribed. The specific universal tools, designated supports, and accommodations approved by Smarter Balanced may undergo change if additional tools, supports, or accommodations are identified for the assessment based on state experience or research findings. The Consortium has established a standing committee, including members from Consortium members and staff, that reviews suggested additional universal tools, designated supports, and accommodations to determine if changes are warranted. Proposed changes to the list of universal tools, designated supports, and accommodations are brought to consortium members for review, input, and vote for approval. Furthermore, states may issue temporary approvals (i.e., one summative assessment administration) for individual, unique student accommodations. It is expected that states will evaluate formal requests for unique accommodations and determine whether the request poses a threat to the measurement of the construct. Upon issuing temporary approval, the petitioning state can send documentation of the approval to the Consortium. The Consortium will consider all state-approved temporary accommodations as part of the annual Consortium accommodations review process. The Consortium will provide to member states a list of the temporary accommodations issued by states that are not Consortium-approved accommodations.

Guidelines for Accessibility for English Language Learners. In addition to the use of Universal Design features, Smarter Balanced has built a framework of accessibility for all students, including English Language Learners (ELLs) that were established in the *Smarter Balanced Guidelines for Accessibility for English Language Learners*. ELLs have not yet acquired complete proficiency in English. For ELLs, the most significant accessibility issue concerns the nature of the language used in the assessments. The use of language that is not fully accessible can be regarded as a source of invalidity that affects the resulting test score interpretations by introducing construct-irrelevant variance. Although there are many validity issues related to the assessment of ELLs, the main threat to validity when assessing content knowledge stems from language factors that are not relevant to the construct of interest. The goal of these ELL guidelines was to minimize factors that are thought to contribute to such construct-irrelevant variance. Adherence to these guidelines helped ensure that, to the greatest extent possible, the Smarter Balanced assessments administered to ELLs measure the intended targets. The ELL *Guidelines* were intended primarily to inform Smarter Balanced assessment developers or other educational practitioners, including content specialists and testing coordinators.

For assessments, an important distinction is between content-related language that is the target of instruction versus language that is not content-related. For example, the use of words with specific technical meaning, such as “slope” when used in algebra or “population” when used in biology, should be used to assess content knowledge for all students. In contrast, greater caution should be exercised when including words that are not directly related to the domain. ELLs may have had cultural and social experiences that

differ from those of other students. Caution should be exercised in assuming that ELLs have the same degree of familiarity with concepts or objects occurring in situational contexts. The recommendation was to use contexts or objects based on classroom or school experiences rather than ones that are based outside of school. For example, in constructing mathematics items, it is preferable to use common school objects, such as books and pencils, rather than objects in the home, such as kitchen appliances, to reduce the potential for construct-irrelevant variance associated with a test item. When the construct of interest includes a language component, the decisions regarding the proper use of language becomes more nuanced. If the construct assessed is the ability to explain a mathematical concept, then the decisions depend on how the construct is defined. If the construct includes the use of specific language skills, such as the ability to explain a concept in an innovative context, then it is appropriate to assess these skills. In ELA\literacy, there is greater uncertainty as to item development approaches that faithfully reflect the construct while avoiding language inaccessible for ELLs. The decisions of what best constitutes an item can rely on the content standards, definition of the construct, and the interpretation of the claims and assessment targets. For example, if interpreting the meanings in a literary text is the skill assessed, then using the original source materials is acceptable. However, the test item itself—as distinct from the passage or stimulus—should be written so that the task presented to a student is clearly defined using accessible language. Since ELLs taking Smarter Balanced content assessments likely have a range of English proficiency skills, it is also important to consider the accessibility needs across the entire spectrum of proficiency. Since ELLs by definition have not attained complete proficiency in English, the major consideration in developing items is ensuring that the language used is as accessible as possible. The use of accessible language does not guarantee that construct-irrelevant variance will be eliminated, but it is the best strategy for helping ensure valid scores for ELLs and for other students as well.

Using clear and accessible language is a key strategy that minimizes construct-irrelevant variance in items. Language that is part of the construct being measured should not be simplified. For non-content-specific text, the language of presentation should be as clear and as simple as is practical. The following guidelines for the use of accessible language were proposed as guidance in the development of test items. This guidance was not intended to violate other principles of good item construction. From the ELL *Guidelines*, some general principles for the use of accessible language were proposed as follows.

- Design test directions to maximize clarity and ones that minimize the potential for confusion.
- Use vocabulary widely accessible to all students, and avoid unfamiliar vocabulary not directly related to the construct (August, Carlo, & Snow, 2005; Bailey, Huang, Shin, Farnsworth, & Butler, 2007).
- Avoid the use of syntax or vocabulary that is above the test’s target grade level (Borgioli, 2008). The test item should be written at a vocabulary level no higher than the target grade level, and preferably at a slightly lower grade level, to ensure that all students understand the task presented (Young, 2008).
- Keep sentence structures as simple as is possible while expressing the intended meaning. In general, ELLs find a series of simpler, shorter sentences to be more accessible than longer, more complex sentences (Pitoniak, Young, Martiniello, King, Buteux, & Ginsburgh, 2009).
- Consider the impact of cognates (words with a common etymological origin) when developing items and false cognates. These are word pairs or phrases that appear to have the same meaning in two or more languages, but do not. Spanish and English share many cognates, and because the large majority of ELLs speak Spanish as their first language (nationally, more than 75%), the presence of

cognates can inadvertently confuse students and alter the skills being assessed by an item.

Examples of false cognates include: billion (the correct Spanish word is mil millones; not billón, which means *trillion*); deception (engaño; not decepción, which means disappointment); large (grande; not largo, which means long); library (biblioteca; not librería, which means bookstore).

- Do not use cultural references or idiomatic expressions (such as “being on the ball”) that are not equally familiar to all students (Bernhardt, 2005).
- Avoid sentence structures that may be confusing or difficult to follow, such as the use of passive voice or sentences with multiple clauses (Abedi & Lord, 2001; Forster & Olbrei, 1973; Schachter, 1983).
- Do not use syntax that may be confusing or ambiguous, such as using negation or double negatives in constructing test items (Abedi, 2006; Cummins, Kintsch, Reusser, & Weimer, 1988).
- Minimize the use of low-frequency, long, or morphologically complex words and long sentences (Abedi, 2006; Abedi, Lord & Plummer, 1995).
- Teachers can use multiple semiotic representations to convey meaning to students in their classrooms. Assessment developers should also consider ways to create questions using multi-semiotic methods so that students can better understand what is being asked (Kopriva, 2010). This might include greater use of graphical, schematic, or other visual representations to supplement information provided in written form.

Fairness as a Lack of Measurement Bias: Differential Item Functioning Analyses

As part of the validity evidence from internal structure, differential item functioning (DIF) analyses were conducted on the Field Test items. This section presents the evidence to support the frameworks’ claims. Chapters 6 and 8 presents the DIF methodology used and results for the Pilot- and Field Test phases. DIF analyses are used to identify those items for which identifiable groups of students (e.g., males, females) with the same underlying level of ability have different probabilities of answering an item correctly or obtaining a given score level. Students are separated into relevant subgroups based on ethnicity, gender, or other demographic characteristics for DIF analyses. Students in each subgroup are then ranked relative to their total test score (conditioning on ability). Students in the focal group (e.g., females) are then compared to students in the reference group (e.g., males) relative to their performance on individual items. It is part of the Smarter Balanced framework to have ongoing study and review of findings to inform iterative, data-driven decisions. These efforts are to ensure that items are not differentially difficult for any group of students.

Test Fairness and Implications for Ongoing Research

There are many features of the Smarter Balanced assessments that support equitable assessment across all groups of students. The assessments are developed using the principles of evidence-centered design and universal test design. Test accommodations are provided for students with disabilities, and language-tools and supports were developed for ELLs. The Work Group for Accessibility and Accommodations and the Consortium developed a set of guidelines to facilitate accessibility to the assessments. In addition to these general accessibility guidelines embedded in the conceptual framework, procedures for item writing and

reviewing and guidelines for creating audio, sign language, and tactile versions of the items were implemented. Smarter Balanced 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, 2009), and a rubric specifically geared towards scoring language complexity (Cook & MacDonald, 2013). In addition, measurement bias was investigated using DIF methods. This evidence underscores the commitment to fair and equitable assessment for all students, regardless of their gender, cultural heritage, disability status, native language, and other characteristics. Irrespective of these proactive development activities designed to promote equitable assessments, further validity evidence that the assessments are fair for all groups of students should be provided. Many of the equity issues are delineated in the most recent version of the NCLB *Peer Review Guidance* (U.S. Department of Education, 2009). To evaluate the degree to which the Smarter Balanced assessments are fulfilling the purpose of valid, reliable, and fair information that is equitable for all students, several types of additional evidence are recommended based on the relevant types listed in the AERA et al. (2014) *Standards*. Validity studies are described here as well as ones that can be addressed in the ongoing research agenda for Smarter Balanced.

Internal Structure. When evaluating the comparability of different variations of a test, such as different language glossaries or accommodated test administrations, validity evidence based on internal structure is the most common approach (Sireci, Han, & Wells, 2008). These studies most often involve multigroup factor analysis (Ercikan & Koh, 2005) or weighted (multigroup) multidimensional scaling, which has also been used for this purpose (see Chapter 5 Pilot Test; Robin, Sireci, & Hambleton, 2003; Sireci & Wells, 2010). Another important source of validity evidence to support equitable assessment is analysis of differential item functioning across test variations and across subgroups of students using differential bundle functioning (Banks, 2013). DIF studies conducted for Smarter Balanced used several criteria to distinguish statistically significant DIF from substantively meaningful DIF (i.e., reflects construct-irrelevant variance). The presence of DIF does not necessarily indicate bias, and therefore, As part of the data review process described in Chapter 3 DIF studies were followed by qualitative analyses that sought to interpret sources of DIF.

Response Processes. Validity evidence based on the relevant subgroups of students were addressed to examine the amount of time it takes different groups of students to respond to items (i.e., item response time) with and without accommodations. Cognitive interviews or think-aloud protocols should be conducted to evaluate the skills measured by items. In addition, specific studies are needed to evaluate accommodations for ELLs or students with disabilities and should be conducted to determine whether the students are using the accommodations and finding them helpful (Duncan, Parant, Chen, Ferrara, Johnson, Oppler, Shieh, 2005).

Relationships with Other Variables. Two types of evidence based on relations to other variables are relevant for validating that the Smarter Balanced assessments are equitable for all subgroups of students (Dorans, 2004). The first type is differential predictive validity studies that evaluate the consistency of the degree to which the assessments predict external criteria across subgroups of students. Zwick and Schlemmer (2004) provided an example of this approach with respect to the differential predictive validity of the SAT for native English speakers and non-native English speakers. These studies are particularly relevant for the “on track” and “college and career readiness” goals of Smarter Balanced. Observational studies using grouping variables could also be conducted using an expected hypothesis of no difference across groups. For

example, by using changes in students' scale scores over time as the dependent variable, comparisons could be made across students from different ethnic groups, socioeconomic status, gender, and other demographic characteristics.

Test Consequences. The analysis of the assessment results can be used to determine if there are differential consequences for various types of students. In describing validity, studies based on testing consequences investigating the effects on instruction, teacher morale, and on students' emotions and behaviors (e.g., dropouts, course-taking patterns) can be conducted. These types of results could also be broken out by subgroup, but more important, the changes in instructional decisions for students should be investigated at the subgroup level. Some important questions might include: Are minority students dropping out of school at higher rates? Are the success rates for remedial programs higher for different types of students?

References

- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219-234.
- Abedi, J., Lord, C., & Plummer, J. (1995). *Language background as a variable in NAEP mathematics performance* (CSE Technical Report 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disability Research and Practice, 20*(1), 50-57.
- Bailey, A. L., Huang, B. H., Shin, H W., Farnsworth, T., & Butler, F. A., (2007) *Developing academic English language proficiency prototypes for 5th grade reading: Psychometric and linguistic profiles of tasks* (CSE Technical Report 727). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Banks, K. (2013). A Synthesis of the Peer-Reviewed Differential Bundle Functioning Research, *Educational Measurement: Issues and Practice, 32*, 43 -55.
- Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*, 133-150.
- Borgioli, G. (2008). Equity for English language learners in mathematics classrooms. *Teaching Children Mathematics, 15*, 185-191.
- Camilli, G. (2006). Test Fairness. In R. L. Brennan (Ed.), *Educational Measurement*, 221-256. Washington, DC: American Council on Education/Praeger.
- Cole, N.S., & Zieky, M. J. (2001). The New Faces of Fairness. *Journal of Educational Measurement. 38*, 4.
- Cook, H.G. & McDonald, R. (2013). Tool to Evaluate Language Complexity of Test Items. Wisconsin Center for Education Research. www.wcer.wisc.edu/publications/.../working_paper_no_2013_05.pdf
- Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*, 405-438.

- 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. (2005). Study of a Dual-language Test Booklet in Eighth-grade Mathematics. *Applied Measurement in Education, 18*, 129-161.
- 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.
- ETS. (2009). *ETS Guidelines for Fairness Review of Assessments*. Princeton, NJ: ETS.
- ETS. (2012). *Smarter Balanced Assessment Consortium: Bias and Sensitivity Guidelines*. Princeton, NJ: ETS.
- Forster, K. I. & Olbrei, I. (1973). Semantic heuristics and syntactic trial. *Cognition, 2*, 319-347.
- Kopriva, R. (2010, September). *Building on student strengths or how to test ELs against challenging math (and science) standards when they don't have the English yet*. Common Core State Standards Implementation Conference.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*. Washington, DC: American Council on Education.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-Centered Assessment Design*. Princeton, NJ: Educational Testing Service.
- Pitoniak, M., Young, J. W., Martiniello, M., King, T., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners*. Princeton, NJ: Educational Testing Service.
- 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.
- Russell, M. (2011b). *Digital Test Delivery: Empowering Accessible Test Design to Increase Test Validity for All Students*. Paper prepared for Arabella Advisors.
- Schachter, P. (1983). *On syntactic categories*. Bloomington, IN: Indiana University Linguistics Club.
- Shafer W., L., & Rivera, C. (2011). Are EL needs being defined appropriately for the next generation of computer-based tests? *AccELLerate, 3(2)*, 12-14.
- 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.

- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M. L., Quenemoen, R. F., & Lazarus, S. S. (2011). *Meeting the Needs of Special Education Students: Recommendations for the Race to the Top Consortia and States*. Paper prepared for Arabella Advisors.
- Young, J., Pitoniak, M. J., King, T. C., & Ayad, E. (2012). *Smarter Balanced Assessment Consortium: Guidelines for Accessibility for English Language Learners*. Available from <http://www.smarterbalanced.org/smarter-balanced-assessments/>
- Young, J. W. (2008, December). Ensuring valid content tests for English language learners. *R&D Connections, No. 8*. Princeton, NJ: Educational Testing Service.
- Zwick, R., & Schlemer, L. (2004). SAT Validity for Linguistic Minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice, 23*, 6-16.
- Zwick, R., & Schlemer, L. (2004). SAT Validity for Linguistic Minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice, 23*, 6-16.