SMARTER BALANCED ASSESSMENT CONSORTIUM:

Bias and Sensitivity Guidelines

June 30, 2022

Table of Contents

INTRODUCTION

PURPOSE

The purpose of the *Smarter Balanced Assessment Consortium Bias and Sensitivity Guidelines* (hereafter referred to as "the *Guidelines*") is to support the process of developing and reviewing Smarter Balanced assessments that 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. The *Guidelines* provide a foundation for diverse item writers and reviewers to address multiple viewpoints when evaluating assessment content for bias and sensitivity.

Consistent with the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), fairness in assessment can be approached by ensuring that test materials are as free as possible from unnecessary barriers to the success of diverse groups of test takers. Those unnecessary barriers can be reduced by following some fundamental guidelines:

- Test items and tasks avoid measuring irrelevant knowledge or skill that advantages or disadvantages particular individuals or subgroups.
- Test content avoids angering, offending, upsetting, or otherwise distracting test takers.
- Test materials are grounded in respectful representation and treatment of all representatives of various cultures (ethnicity, race, nationality, language, gender, disability, neurodiversity, age, sexual orientation, place of origin, beliefs, education, professional experience, communication style, and other cultural characteristics).

This document describes in detail how to follow these guidelines for the Smarter Balanced assessments of the state standards in English Language Arts (ELA) and mathematics. Some aspects of the *Guidelines*, found in corresponding sections of this document, might not be appropriate for tests of specific subjects such as Biology or Psychology.

USES

The intended use of the *Guidelines* is in the development of the Smarter Balanced Assessments, particularly in item writing and reviewing, and the primary intended audience includes educators and other professionals involved in these processes. This document describes the rules agreed upon by the Smarter Balanced Assessment Consortium member states and territories for achieving fairness in test content and reducing subjectivity when evaluating test items for fairness. Only items that are in compliance with the *Guidelines* will be included in the Smarter Balanced assessments. The *Guidelines* will help ensure that the test content is fair for test takers, as well as acceptable to the many decision makers and constituent groups within the Smarter Balanced member states and territories. Although many people think of bias and sensitivity guidelines as applying primarily to the review of test items after they have been written, fairness must be considered in all phases of test development, design, and use. This process includes the stages of item writing and reviewing, adding accessibility resources to items grounded in these *Guidelines,* reviewing post-field test data with the *Guidelines* in mind, and reviewing operational items for possible emerging issues related to bias and sensitivity.

Use of the *Guidelines* will help the Smarter Balanced assessments comply with Standard 3.2 of the *Standards for Educational and Psychological Testing*:

> Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics (AERA, APA, & NCME, 2014, p. 64).

LINK TO EVIDENCE-CENTERED DESIGN

The Smarter Balanced assessments are developed using the principles of Evidence-Centered Design (ECD). Three basic elements of ECD are 1) stating the claims to be made about test takers, 2) deciding what evidence is required to support the claims, and 3) administering test items that provide the required evidence (Mislevy, Steinberg, & Almond, 1999).

ECD provides a chain of evidence-based reasoning linking test performance to the claims to be made about test takers. Fair assessments are essential to the use of ECD. If test items are not fair, then the evidence they provide is not an accurate representation of the skills and knowledge of the subgroup for which the items are biased. Under these circumstances, the claims cannot be equally well-supported for all test takers. Therefore, appropriate use of the *Guidelines* helps to ensure that the evidence provided by the items means the same thing for participating test takers and allows ECD to work as intended.

DEFINING VALIDITY, BIAS, SENSITIVITY, AND FAIRNESS

VALIDITY

To define "fairness" and "bias" for the purposes of the *Guidelines*, it is necessary to understand the meaning of "validity." Validity is the extent to which the inferences and actions made on the basis of test scores are appropriate and backed by evidence (Messick, 1989). More simply, validity can be thought of as the extent to which test scores accurately reflect the relevant knowledge and skills of test takers. For Smarter Balanced assessments, the relevant knowledge and skills are defined by state standards.

FAIRNESS

When addressing issues of fair testing and attending to bias and sensitivity considerations, the *Standards for Educational and Psychological Testing* indicate "a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct. To the degree possible, characteristics of all individuals in the attended test population, including those associated with race, ethnicity, gender, age, socioeconomic status, or linguistic or cultural background, must be considered throughout all stages of development, administration, scoring, interpretation, and use so that barriers to fair assessment can be reduced" (AERA, APA, & NCME, 2014, p. 50). "Sensitivity" is used to refer to an awareness of the need to avoid bias in assessment. In common usage, test and item review for bias and sensitivity helps

ensure that the test items and stimuli (or reading passages) are fair for various groups of test takers and takes into account overall accessibility of assessment materials through the lens of diversity, equity, and inclusion.

"Fairness" is a more difficult word to define because, as indicated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 49), "*fairness* is used in many different ways and has no single technical meaning." Elliott (2016) defines fairness in assessment as "the identification of opportunity structures created through maximum construct representation under conditions of constraint—and the toleration of constraint only to the extent to which benefits are realized for the least advantaged—expressed in terms of its tradition, boundary, order and foundation" (p. 1). A useful definition of fairness for the purposes of the *Guidelines* is the extent to which the test scores are valid for different groups of test takers. For example, a math item may contain difficult language unrelated to mathematics. If the language interferes equally with all test takers, validity will be reduced for all test takers, but the item will not necessarily be unfair. If, however, the language is a bigger barrier for students who are not native speakers of English than for other students, then the item will be unfair.

Even if the 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 is intended to measure a test taker's English language reading comprehension ability, score differences between groups based on real differences in comprehension of English would be valid and, therefore, fair.

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, both in terms of the item's meaning and presentation. 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 and, thus, were unable to use their attention and energy to focus on the content of the test. The notion of fairness in assessment content is continuously evolving (Sireci & Randall, 2021), and these *Guidelines* will be updated correspondingly.

QUANTITATIVE AND QUALITATIVE EVALUATIONS OF FAIRNESS

QUALITATIVE EVALUATIONS

No quantitative indicator of fairness can replace human judgment from multiple participants in evaluations of fairness. Issues that may affect fairness are often too subtle to be captured by any statistic. Therefore, the primary strategy against the inclusion of unfair materials in the Smarter Balanced assessments is the judgment of trained test developers who should represent diverse cultural backgrounds and who follow the *Guidelines.* Test developers may, however, miss potential fairness issues that sophisticated statistical analyses later find (Bond, 1993). Therefore, both qualitative and quantitative evaluations of fairness are required. There must be a balance between striving to ensure fairness and the ability to measure the full range of state standards with authentic and interesting materials. Proper use of the *Guidelines* will help maintain that balance.

QUANTITATIVE EVALUATIONS

In addition to judgmental reviews for fairness, items in the Smarter Balanced assessments receive a quantitative check for fairness. Items in the assessments are field tested with a representative sample of students to see how well the items work before they are used to evaluate students. At that stage, a statistic called Differential Item Functioning (DIF) is used as a statistical indicator of fairness. DIF studies are required by Cluster 2, Standards 3.6-3.8 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, pp. 65-66).

It is important to note that there exist limitations of DIF analyses when it comes to identifying biases, and that other research methods need to be investigated. Merely calculating differences in test item difficulty is not a useful indicator of fairness, because between-group differences for test takers may be valid. That is, groups of test-takers may actually differ on the relevant knowledge that an item is supposed to measure. Due to lack of opportunity to learn, etc., group differences may be indicators of many factors that contribute to the disparity in knowledge, but not necessarily that the test item itself is unfair. DIF uses a variety of statistical analyses based on the straightforward concept that people who have the same knowledge about a subject, should perform similarly on test items about that subject, regardless of any within- or between-group differences, such as gender or race. Test scores often determine which test takers have the same knowledge about a subject, and people with the same or very similar scores are considered to be "matched." Significant differences in test item difficulties or matched people in different groups, result in higher DIF statistical values.

DIF alone, however, is not proof of bias. No test is perfect. Therefore, no matching of test takers on the basis of test scores can be perfect. A fair item may show DIF merely because the test scores have not matched people well regarding particular knowledge or skills that are validly measured by the item (see Dorans, 1989; Holland & Thayer, 1988; and Zieky, 1993 for more information about DIF and its uses in test development). It is also important to keep in mind that there are other aspects that may contribute to statistically significant differences that have implications for validity inferences; for instance, students may not be familiar with a specific item format.

COMBINATIONS OF EVALUATIONS

Neither DIF nor any other statistic can be considered proof that an item is either fair or biased, but appropriate statistics can help identify any potentially unfair items. Using a combination of qualitative and quantitative analyses of the performances of matched test takers in different groups is the best method available to help ensure the fairness of the Smarter Balanced assessments. Smarter Balanced is committed to revising the *Guidelines* on a regular basis and to adjusting its qualitative and quantitative analyses accordingly.

ISSUES OFTEN CONFUSED WITH FAIRNESS

TEST ITEM DIFFICULTY

Test item difficulty should not be confused with fairness. A difficult item is not necessarily unfair if the sources of difficulty are valid. For example, state standards call for students to read primary source documents from the United States history. Some of these documents are difficult to read, and valid items will appropriately reflect this difficulty.

OVEREXTENDING GUIDELINES

Reviewers should avoid overextending the *Guidelines* to contrive situations in which an innocuous topic is judged to be unfair. That practice inappropriately limits test content because any topic can be judged to be potentially upsetting in some set of circumstances for some test takers. For example, a reviewer might say that an innocuous depiction of a mother with her child might upset a test taker who has lost a parent. A topic that is upsetting in general is probably unfair, but an innocuous topic that might possibly be upsetting under some particular set of circumstances is not necessarily unfair.

CONTENT AND LANGUAGE THAT ARE ACCEPTABLE TO INCLUDE IN SMARTER BALANCED ASSESSMENTS

ELIGIBLE CONTENT

With respect to the validity and fairness of the Smarter Balanced assessments, any content that is required by state standards, consistent with the Consortium's item specifications guidelines, and reviewed through the processes described by the Consortium's item review procedures may be included in the Smarter Balanced assessments. Additionally, state laws and state policies in one or more of the member states or territories may also affect the assessment content that member states or territories present to their students. For example, a state law may require the inclusion of content based on the achievements of specific groups. Any content required by state law or state policy may be administered during the same test event as a Smarter Balanced assessment but needs to conform to the Consortium's policies regarding the addition of state-specific content. These policies may include, at a minimum, the ability to derive a score that excludes the additional state content such that a comparable score may be reported for all students taking a Smarter Balanced assessment.

EXPOSURE TO INFORMATION

Stimuli (reading passages) for ELA items have to be about some topic. Mathematics problems are often placed in real-world contexts. Often state standards do not include all of the content areas (outside of ELA and mathematics) from which topics and contexts must be drawn. Which topics and contexts are fair to

include in the Smarter Balanced assessments? One fairness concern is that students differ in exposure to information through their life experiences outside of school. For example, some students experience snow every winter, while some students have never experienced snow. Some students swim in the ocean every summer, while some students have never seen an ocean. Some students live in houses, some live in apartments, some live in mobile homes, and some are unhoused.

Even though curricula differ across and within Smarter Balanced members, the concepts to which students are exposed in school tend to be much more similar when compared to students' life experiences outside of school. If students have become familiar with concepts through classroom exposure, the use of those concepts as topics and contexts in test materials is fair, even if some students have not been exposed to the concepts through their life experiences. For example, a student in grade 4 should know what an ocean is through classroom exposure to the concept, even if that student has never actually seen an ocean. A student does not have to live in a house to know what a house is, if there has been classroom exposure to the term. Similarly, a student does not have to be able to run in a race to know what a race is. Mention of snow does not make an item unacceptable for students living in warmer parts of the country if they have been exposed to the concept of snow in school.

INFORMATION IN THE STIMULUS

A major purpose of reading is to learn about new things. Therefore, it is necessary to include material that may be unfamiliar to students if the information necessary to answer the items is included in the tested material. In this case, it is helpful to employ a *window-mirror* approach—the student is presented with sufficient content through the *window* of the stimulus or item so that the student is able to *mirror* that content back and respond to the test item meaningfully. For example, it is fair to test the ability of a student who has never been to a desert to comprehend an appropriate stimulus about a desert, as long as it includes the information needed to answer the test item.

ACCESSIBILITY GUIDELINES

The Consortium uses the principles of Universal Design (Johnstone, Altman, & Thurlow, 2006) that promote inclusivity of diverse test populations; precisely defined constructs; accessible, non-biased items; tests that are amenable to accommodations; simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility. These principles help reduce visual, auditory, cognitive, communicative, physical, and other barriers to valid measurement, but some accommodations for some students with disabilities or certain test content/formats are still necessary. Barriers related to physical abilities occur if test takers have difficulty seeing or hearing the test materials or have physical difficulty responding to the test under standard conditions (e.g., manipulating a computer mouse). Other barriers, such as unnecessarily complex text or unfamiliar phrases, may also affect students' ability to show what they know and can do. Barriers to fairness are detailed in this paper as well.

ACCESSIBILITY FOR ALL STUDENTS

Accessibility resources that include individualizable embedded and non-embedded universal tools, designated supports, and accommodations are developed with diverse students in mind because accessibility needs and preferences are unique and can be customized for all participating general education students, not

only those who may have an English learner or disability status. Assessment content needs to be developed with these considerations in mind so that the universal design approach is infused in all stages of item and stimulus development.

See the *Smarter Balanced Assessment Consortium Usability, Accessibility, and Accommodations Guidelines* (2021*)* for information about how to ensure that the Smarter Balanced assessments are as accessible as possible for participating students, including English learners, students with disabilities, English learners with disabilities, and other general education students.

STUDENTS WITH DISABILITIES

A similar consideration arises for students with disabilities. It is important to consider whether it is fair to include material about the visual arts or music for students who cannot experience them directly. There may be an undue burden in defining what those experiences might infer, so that students with disabilities need to spend more time and energy figuring out the context of the item than their peers. It may also be unfair to include stimuli about physical activities for students who cannot participate in them. As noted above, it is acceptable to include material that may be unfamiliar to some students based on their life experiences, as long as the information necessary to answer the items is included in the stimuli or is part of the information expected from classroom exposure. For students with certain disabilities, it is necessary to add the provision that the information necessary to answer the items does not need to be obtained through direct, personal experience. For example, a high school student who is deaf could fairly be expected to know what a guitar is, but could not fairly be expected to know what a guitar sounds like. Assertions that using all of one's senses provides an advantage in life, such as "lucky enough to see," "listening to music is the best way to relax," "playing an instrument results in better grades," or "aspiring careers that prioritize physical abilities" should not be included in assessment content.

ENGLISH LEARNERS

Unnecessarily complex language can be a source of unfairness when the language itself is not the focus of measurement. This is particularly true for English learners. The language in mathematics items should not be a barrier to a correct answer for students who could do the required mathematics. For mathematics assessments, therefore, nonmathematical language should be clear and simple and targeted no higher than the grade level below the tested grade level. The focus of ELA assessments is on the use of language, although unnecessary complexities need to be avoided in this content area as well. Valid assessment of ELA state standards requires the use of language targeted at the tested grade level. In general, the clearest language consistent with validity should be used when the language itself is not being tested. Special attention should be paid to cultural issues to ensure that cultural misnomers, such as treating Africa as a country or using *American* where *U.S. American* would be more appropriate, are not included in assessments. It is also important to be mindful of the barriers to fairness described below and avoid cultural stereotyping and cultural tokenism that occur when aspects of culture are acknowledged inadequately or simply because someone is trying to "check a box." Finally, it is important to prioritize ethnorelative (culture-general) content and either avoid or sufficiently describe ethnocentric (culture-specific) content. For instance, if the item is based on the rules of baseball, those rules need to be clearly outlined in the item. For more detail on appropriate language for English learners, see the *Smarter Balanced Assessment Consortium Accessibility Guidelines for English Language Learners* (2012*)*.

BARRIERS TO FAIRNESS

Fairness reviews are intended to look at each item or stimulus through different lenses and remove barriers to valid measurement that may affect different groups of test takers in different ways. There are two major types of barriers related to fairness—those stemming from irrelevant knowledge and those stemming from stress.

BARRIERS STEMMING FROM IRRELEVANT KNOWLEDGE

Barriers stemming from irrelevant knowledge occur when uncommon information—not reasonably expected of some group(s) of students and not related to state standards—is required to answer a test item. For example, assuming that a student knows what a "foyer" is would be unfair because the term 1) is more likely to be known by some groups of students than by other groups of students, 2) is not required by state standards, and 3) is not likely to have been routinely used in the classroom.

BARRIERS STEMMING FROM STRESS

Barriers stemming from emotional reactions may occur if language or images cause strong emotional reactions among members of some groups of test takers and those reactions potentially interfere with test performance. For example, if a passage advocates for one position on a controversial issue such as gun control, students who are strong supporters of the opposite position may be disadvantaged by having to put their beliefs aside to respond correctly to items associated with this passage.

When depicting a dangerous situation, it is important to avoid content related to emotional or psychological trauma students may have experienced as a result of stressful events that may have interfered with their sense of security. For example, content related to pandemics or natural disasters must undergo thorough reviews to ensure that those topics do not provoke any feelings that can potentially traumatize students.

Even if student performance is not directly affected, the presence of offensive, inflammatory, controversial, upsetting, and disrespectful material in tests will lower the confidence of students, parents, politicians, educators, and other community members in the fairness of the test.

AVOIDANCE OF IRRELEVANT KNOWLEDGE

It is necessary to avoid unfair barriers to success based on group differences in knowledge unrelated to the purpose of the test. Requiring specialized irrelevant knowledge to answer a test item is unfair. For example, it is unfair to require prior knowledge of the number of people on a hockey team to answer an item in the Smarter Balanced assessments because students who know the relevant content may not have the irrelevant knowledge of hockey needed to answer the item. (Note that testing specialized knowledge is appropriate when that knowledge is relevant to the purpose of the test. Requiring knowledge about the number of people on a hockey team may be perfectly appropriate on other content area tests.) This guideline prohibits the testing of specialized knowledge when that knowledge is not relevant to the purpose of the test. Specialized knowledge that is explained in the stimulus material or can be inferred by contextual clues is

acceptable, if understanding the explanation or making inferences from the explanation is what is being assessed.

The following categories are common sources of specialized irrelevant knowledge in some large-scale assessment content. Other barriers such as unnecessarily complex text or unfamiliar phrases may also affect students' ability to show what they know and can do. Familiarity with these topics and other similarly specialized knowledge—when unrelated to the purpose of the test—should not be required to respond to items, unless the necessary information is provided in the stimulus material. It is important to remember that some standards may require including terms belonging to these categories, but it is important to ensure that their use adheres to the guidelines described here.

REGIONALISMS

Avoid requiring knowledge of words and phenomena limited to a region or certain regions of the country and words that carry different meanings in different regions, e.g., "hero" for "sandwich," "snow days" at school, "tonic" or "pop" for "soda," "muffler" as an article of clothing, and "bubbler" for "water fountain."

RELIGION

Required knowledge about any particular religion, as well as content that privileges a particular religion or system of beliefs over others, is best avoided. For example, to say that something is "as colorful as an Easter egg" may be an unfamiliar comparison for some students.

OCCUPATIONAL AND TECHNICAL INFORMATION

Avoid requiring knowledge of specialized information and terminology—not related to the purpose of the test—that is associated with a particular occupation or field of knowledge such as agriculture, law, mechanics, military, science, sports, technology, transportation, or weapons. For example, avoid requiring irrelevant knowledge about the purpose of a silo, the less common names for tools, the chain of command in military organizations, the functions of parts of weapons, the scoring systems or rules of play in various sports, the uses of a flange, or the meaning of "lumen."

The point at which words become overly specialized is a matter of judgment and will vary by grade level. Therefore, content experts at the tested grade level are best equipped to judge the appropriateness of words associated with a particular field of knowledge.

SLANG

Avoid highly contextual slang terminology that may not be familiar to many test takers if it is not appropriately defined in the assessment content. Examples include "frenemy," "brb," and "hangry." Some ELA literary texts may include instances of slang language as measured by the corresponding standards.

ACADEMIC LANGUAGE

In general, academic language is expected in assessment content. However, overuse of academic terms in construct-irrelevant parts of an item may interfere with assessment validity. For example, instructions or directions that use “register your answer” instead of “enter your answer,” demonstrate this point.

FIGURES OF SPEECH

Avoid requiring understanding of figures of speech, such as idioms, metaphors, epithets, hyperboles, similes, and others (e.g., spill the beans, hit the hay, fly in the ointment, flash in the pan, drowning in paperwork, dynamic duo, bored to tears, as light as a feather), unless understanding a figure of speech is needed to respond to ELA items in state standards. In such instances, determining the meaning and impact of figures of speech, including idioms, through context clues is part of the eligible content for both literary and informational reading passages.

COMMERCIAL BRAND NAMES AND TECHNOLOGY

Avoid using the names of commercial brands (e.g., Android, iPhone), companies (e.g., Amazon, Facebook), and other business entities so as not to give the impression that the Consortium endorses or promotes a particular brand or product. Usage of commercial brands and product names can also suggest disparities in social class regarding assumptions about who might have access to these consumer goods. Finally, technology changes rapidly and the use of commercial brands or specific technology may cause the test content to become irrelevant in future years (e.g., floppy disk, Myspace).

GENDER NEUTRALITY

Avoid using *he* as a universal pronoun; likewise, avoid using binary alternatives such as “he/she,” “he or she,” or “(s)he.” When possible/appropriate, use the student’s name or the term “student” itself and use assigned pronouns consistently.

TOPICS TO AVOID

Certain topics are extremely controversial, upsetting, inflammatory, and often judged by parents and communities to be inappropriate for children. Such topics should be excluded from the Smarter Balanced assessments unless they are required to measure the state standards. The goal is to avoid material that test takers may find extremely upsetting, as strong negative emotions can potentially interfere with test performance. While some of these subjects may be discussed in classrooms and can generate rich discussions, test takers do not have the opportunity to discuss subjects that may upset them during the test. It is best not to include materials that may cause strong negative emotions such as anger, disgust, fear, hatred, or sadness.

The following list is intended to indicate the nature of topics that should be excluded from Smarter Balanced assessments, but the list is not exhaustive. Current and emerging events may add topics that are so problematic that they should be excluded from the assessments. Topics to be avoided include, for example:

- abortion
- abuse of people or animals
- contraception
- deportation of immigrants
- experimentation using animals that is dangerous or painful
- killing of animals for sport
- the occult, witches, ghosts, or vampires
- rape
- sexual behavior or sexual innuendo
- suicide
- torture

TOPICS TO BE TREATED WITH CARE

There are also a number of topics that need to be treated with care, particularly those topics that pertain to sensitive or controversial issues. Although such topics can be included for historical purposes or with appropriate contextualization or framing, special attention needs to be paid to how these issues are presented and to how fair they are to test takers. Examples of such topics include:

- climate change caused by human behavior
- current or recent partisan political issues, ethnic conflicts, religious disputes, and other controversial current events
- euthanasia
- gun control
- pandemics, infectious diseases, and vaccines
- prayer in school

Other sensitive but less upsetting topics may be included in Smarter Balanced assessments. Such topics must, however, be treated carefully to minimize potential fairness issues. Guidelines that forbid a topic are easy to apply. Guidelines that require treating a topic with care are more difficult to apply because different people will have different opinions about what is acceptable. Consideration of a topic should include the degree to which a student's reaction to a topic might hinder the student's ability to demonstrate fully what they know and can do in relation to the item or stimulus.

When making judgments about the suitability of materials on topics such as those listed below, it is important to keep in mind that the Smarter Balanced assessments must not only be fair and valid for test takers, they must also appear to be fair and valid in the opinions of various constituencies within the Consortium. It is counterproductive to use test materials that various groups within the Consortium will consider inappropriate for their children.

ACCIDENTS AND NATURAL DISASTERS

Mention of these topics or general, objective discussions may be acceptable, but avoid a focus on suffering, destruction, loss of life, loss of homes, or graphic, gruesome details that may upset or frighten students.

ADVOCACY

The Smarter Balanced assessments should not support one side of a controversial issue. Avoid advocacy when possible because test takers with opposite views may be disadvantaged. If, however, advocacy is required to measure a state standard, indicate that the material does not necessarily represent the views of the Consortium. Additionally, avoid advocating for or against a political party unless doing so is important to measure a state standard.

ALCOHOL, TOBACCO, AND ILLEGAL DRUGS

It is best to avoid depictions of people using alcohol, tobacco, and illegal drugs so as to avoid giving any impression that these substances are approved. Do not depict use of these substances as pleasurable, alluring, or as signs of sophistication and maturity. Warnings against the use of these substances may be acceptable for students in middle or high school.

ANIMALS THAT ARE FRIGHTENING TO CHILDREN

Younger students are more likely to be upset by certain dangerous animals than are older students. Avoid depictions of spiders and poisonous snakes because it can cause problems for some children. Objective depictions of a food chain or nonthreatening descriptions of animals are acceptable, but avoid depicting predators engaged in violent, threatening behavior. For example, a discussion of how members of a wolf pack interact with each other is likely to be acceptable. Avoid a depiction of a wolf ripping the entrails from a fawn or attacking a child.

BIOGRAPHICAL MATERIALS

Take care in selecting biographical materials. Some biographical materials may be controversial because different groups of people may view the individuals depicted very differently. It is important to keep in mind that one group's heroic freedom fighter could be another group's cowardly terrorist. A possible concern with the use of biographical material about living people is that persons who are widely admired at the time they are included in test materials may become involved in a highly publicized scandal before the test is administered.

COLONIALISM

Western colonialism is a political-economic phenomenon where various European nations explored, conquered, settled, and exploited large areas of the world. Colonialism is considered acts of genocide of indigenous peoples that resulted in the mass destruction of entire communities of indigenous peoples. American Indian and Alaska Native people face pervasive stereotypes, misconceptions, and omissions about their history and identity in texts. Historical text selections should be reviewed for aspects of colonialism.

Many historical events such as the California Missions, Gold Rush Allotment Period, American Indian Boarding Schools, and Self-determination are considered acts of genocide, cultural erasure, and historical trauma for American Indian and Alaska Native people.

CURRENT EVENTS AND CONTEMPORARY PERSONS

Content depicting events related to or people involved in negative contexts, such as criminal activity, violations of human rights, or racism, should be treated with care. Special attention should be paid to the extent to which controversial content is detailed and whether the content is generally appropriate.

DANCING

Allow all forms of dance except couples social dancing, which is the type most likely to draw criticism from some groups.

DANGEROUS ACTIVITIES

Avoid modeling behaviors that are inherently dangerous or making dangerous behaviors appear to be attractive, fun, glamorous, or something to be emulated. Particularly for younger children, avoid showing potentially dangerous behavior such as running away from home, going with strangers, or using dangerous tools or weapons without supervision. Common actions that are dangerous if done improperly (such as crossing the street, riding a bicycle, hiking, or swimming) are acceptable if depicted as being done properly. It is not acceptable to describe dangerous substances or devices such as weapons, poisons, or explosives in ways that make them appear attractive or safe.

DEATH AND DYING

Detailed depictions of the death of parents, siblings, contemporaries, and family pets should be avoided unless necessary to measure a state standard. It is acceptable to mention death (e.g., Rosa Parks died in 2005), but it is not acceptable to depict gruesome details.

ENSLAVEMENT OF PEOPLE

Special attention needs to be paid to how this topic is presented. Language has shifted from using the word “slave” (or “slaves”) when referring to enslaved Africans, to preserve the humanity of the people stolen from their countries of origin and sold into slavery in the Americas. This topic may be included in assessments as it is important for the measurement of state standards. An anti-racist approach to assessment requires developers to provide complete and accurate historical perspectives that go beyond celebrating Whiteness. Any treatment of the narrative(s) of enslaved peoples should represent the truth of those people (and note that White people should never be the arbitrators of that particular truth). The legacy of the complete history of enslavement permeates every aspect of American society and should be addressed in assessments across multiple content areas.

EVICTIONS AND UNHOUSED CONDITIONS

Discussion of evictions and unhoused conditions may be particularly upsetting to students who have direct experience with evictions and being unhoused or fear experiencing them in the future. These topics must be treated factually and with care. Avoid the emotional discussion of these topics, including aspects that focus on anguish and distress.

EVOLUTION

Because it is highly controversial for some people, the topic of evolution of human beings or similarity of human beings to other primates should be treated with care. Evolution within a species (such as evolution of bacteria to withstand antibiotics) is much less problematic and could be allowed if treated with care. Fossils and the age of Earth are acceptable if not linked to evolution of human beings. (In tests intended to measure knowledge of science, any aspect of evolution required to measure this construct is acceptable.)

FAMILY PROBLEMS

Avoid upsetting test takers with detailed descriptions of serious family problems such as the loss of a job, loss of a home, divorce, detainment, incarceration, or serious illness of a parent or sibling, except as needed in historical or literary materials to measure state standards.

GAMBLING

Instruments used for gambling such as playing cards and dice may be used as required in math problems. However, do not assume that all students will be familiar with them and that all students will know, for example, the number of cards in a deck or the maximum number obtainable on a pair of dice. Avoid depictions of people gambling for fun or profit.

HARMFUL, CRIMINAL, OR INAPPROPRIATE BEHAVIORS

Avoid modeling inappropriate or bad behavior for students so as not to upset anyone who may have been the victims of such behavior by others. Examples of harmful, criminal, or inappropriate behaviors include bullying, cheating, truancy, joining gangs, fighting, lying, and stealing. It is particularly important to avoid making such behavior appear to be attractive, fun, glamorous, sophisticated, or something to be emulated.

HISTORICAL FIGURES, EVENTS, AND PLACES

Narratives related to historical figures that explicitly or implicitly point to those figures' involvement in negative contexts such as criminal activity, for example, should be treated with care. This is also applicable to historical events and places and warrants additional attention to the extent to which those events or places detail controversial content and appropriateness of the content in general. Narratives and counternarratives of contemporary marginalized persons, including their contemporary and historical mistreatment (medical experimentation, internment camp practices, segregation, etc.) should be presented to disrupt widely held racist logic about these communities and their experiences.

HOLIDAYS AND BIRTHDAYS

Mentioning holidays and birthdays is acceptable as long as all of the information necessary to answer items on these topics is included in the stimulus material. Avoid use of religious materials and extended discussion of religious holidays and birthdays. Not all test takers celebrate birthdays, and not all test takers will be familiar with every religious or quasi-religious holiday (e.g., Halloween).

IMMIGRATION

Immigration must be treated factually and objectively if the inclusion of the topic is important to measure a state standard. Test takers or their families may have experienced or are experiencing traumatic events such as detainment or threat of deportation relating to immigration, therefore, this topic needs to be treated with care.

LUXURIES

Test materials do not have to be limited to what might be accessible to the least affluent families. However, avoid elitism and the impression that ordinary people are excluded from the test materials. Avoid discussion of luxuries such as servants, mansions, summer houses, expensive vacations, and yachts unless needed to measure state standards in literary or historical materials. Avoid more common luxuries such as ski trips and private tennis lessons. Avoid depicting expenditures that most people would consider excessive. For example, in a math item, do not have a person purchase three suits at $150 per suit.

MEDICINES

Treatments for serious illnesses may be upsetting to some students and should be avoided. Do not model the use of drugs, even prescription drugs, as a way to solve problems. Some groups are opposed to medical treatment, so it is best to avoid the topic. Avoid topics related to diet supplements.

MENTAL HEALTH

Mental health is a sensitive topic and needs to be treated with care. Avoid stigmatizing mental health issues and using terms often associated with mental health in evaluative ways. For example, pay close attention to the use of such terms as “crazy,” “unhinged,” “insane.” Avoid content that emphasizes advantages or privileges of people who are described as not having challenges that are associated with mental health, such as not having depression or anxiety.

PERSONAL QUESTIONS

Items must not invade the privacy of students by asking them to divulge personal or family issues such as religion, political preference, or antisocial or criminal behavior. For example, do not use an item that asks test takers to describe a time when they were caught doing something wrong. It is best to avoid constructed-response items that require students to reveal how they would act in situations contrary to their beliefs about appropriate behavior.

PHYSICAL APPEARANCE AND ATTRIBUTES

Avoid upsetting children by depicting their heights, weights, or other physical attributes with negative connotations. A wide range of body types should be represented in any written, oral, or visual material, but avoid stereotypes and negative depictions of body shapes and other physical or psychological characteristics.

PREGNANCY OF HUMAN BEINGS

Assessment content should not include topics that portray pregnancy as inappropriate, shameful, or wrong or imply that pregnant people should isolate themselves during the pregnancy period. Students should see representations of pregnant women in various life settings, e.g., running a business meeting, standing in line in a coffee shop, speaking with friends in a park.

RELIGION

Religion is a source of information that is not common or accessible to all students. In Smarter Balanced assessments, religion is a topic best treated with great care. Some people will see even an objective description of a religion as proselytizing. However, it is acceptable to mention religion. For example, noting that Buddhism is one of the main religions in Singapore is acceptable. Going into detail about the practices of adherents of Buddhism is not acceptable. In particular, avoid praising or criticizing the practices of a religion. Also avoid references to God, euphemisms for God, or creationism except in historical or literary documents important for the measurement of state standards.

SERIOUS ILLNESSES

Serious illnesses include mental as well as physical illnesses. Illnesses that primarily affect certain groups, such as some genetic diseases, may be particularly problematic. Mention of serious illnesses may be acceptable, but avoid focusing on suffering or on graphic, gruesome details that may be upsetting to students. Ensure that information about illnesses mentioned is accurate and current. For example, when talking about diet and diabetes, it is important to note the differences between Type I and Type II diabetes.

SOCIAL MEDIA

The topic of social media needs to be treated with care when it comes to depictions of negative social media effects on youth mental health and behaviors such as cyberbullying, harassment online, and social exclusion. Also, special attention needs to be paid to issues related to gaining significant earnings by being a social media influencer or promoting self-worth (e.g., the number of views and followers).

TERRORISM, WARS, VIOLENCE, AND SUFFERING

These topics may be included in historical or literary documents if important to measure state standards. Avoid focusing on graphic, upsetting, or frightening aspects of these topics.

UNHEALTHY BEHAVIOR

The goal is to avoid modeling unhealthy behavior by showing excessive consumption of food or glorifying a lack of exercising. However, it is acceptable to mention eating a cookie, for example, or to use the sharing of a pie to illustrate a fraction.

WESTERN IMPERIALISM

These topics focus on the domination of countries in Europe over countries in Africa, Australasia, and the Americas. Historical aspects of colonial domination that took the form of political, economic, sociocultural, and physical control over people, resources, and labor in these regions should be presented critically. Many western imperialist ideas and structures remain in the former colonies, and the impact on the descendants of the colonized (including those subjects to enslavement and genocide), as well as the colonizers, needs to be allotted special attention.

AVOIDANCE OF STEREOTYPES

Materials in Smarter Balanced assessments should not reinforce stereotypes and should avoid them in general. It is acceptable to depict gender conforming behavior (e.g., a woman caring for children), but gender conforming behaviors must be balanced by depictions of gender nonconforming and/or fluidity to avoid reinforcing stereotypes (e.g., a man caring for children). For adaptive tests (assembled by a computer as they are administered to a student), balance is best handled at the level of the item pool. To help ensure that the item pool is balanced, item writers should produce items showing gender nonconforming and/or fluid behaviors whenever they produce items showing gender conforming behaviors that could be considered stereotypes.

The following types are particularly problematic:

STEREOTYPED LANGUAGE

Some stereotyped language may be acceptable in literary or historical material important for the measurement of a state standard, even if it uses outdated terms and nonparallel language for different genders. In general, avoid phrases such as “man-sized job” or “Dutch uncle.” Language that uses different terms for the same characteristic in men and women is not acceptable. For example, it is not appropriate to label a man as “forceful” or “assertive” and a woman as “pushy” or “controlling” for exhibiting the same behavior. Language that assumes all members of a profession are one gender is unacceptable (e.g., use “sales representative” instead of “salesman,” “firefighter” instead of “fireman,” “mail carrier” instead of “mailman”). Some stereotyped language may be acceptable in literary or historical material important for the measurement of a state standard.

STEREOTYPED IMAGES

Avoid stereotyped images. For example, do not show all girls in frilly dresses and all boys in jeans. Do not show all White men in suits and ties and all Black men dressed as laborers. If it is impossible to show diversity in a single image, diversity should be shown across images.

STEREOTYPED SOCIAL/OCCUPATIONAL ROLES

Materials in Smarter Balanced assessments should ensure that individuals from different background characteristics are represented at different occupational roles with varying levels of power. There should be a mix of genders, races, and other cultural characteristics shown in any social or occupational role. For example, ensure that positions such as doctors, supervisors/bosses, and CEOs include people from different racial/ethnic backgrounds and genders. Do not depict all male doctors with all female nurses. Diversity should be shown across items, if it is impossible to show diversity in a single item.

STEREOTYPED BEHAVIORS AND CHARACTERISTICS

Materials in Smarter Balanced assessments reflect the diversity of the human experience in terms of gender expression, sexual orientation, and racial, ethnic, and/or national identities. Test items and materials should not treat members of any group as though they all share the same characteristics, traits, or lived experience. Do not portray any such group as more (or less) lazy, immoral, primitive, ignorant, prone to crime, gullible, violent, miserly, arrogant, or dirty than any other such group. For example, do not depict all Native American people as close to nature, all Asian American students as smart, or all Latinx/Latine people as solving problems incorrectly and asking test takers to identify those mistakes.

TOPICS TO PRIORITIZE

TOPICS REQUIRED BY STATE STANDARDS

Prioritize topics that promote students' ability to demonstrate independence, respond to various audiences and tasks, analyze evidence, understand other perspectives, as well as other aspects included in state standards.

TOPICS ADDRESSED IN CLASSROOM INSTRUCTION

Include relevant and authentic topics that all students, including students of various cultural representations and classifications (e.g., English learner status, disability category, socioeconomic status, other general education students), have access to in classroom instruction. Examples include topics related to classroom activities, extracurricular activities, and other topics related to students' lives.

CREATIVITY

It is important to prioritize topics that foster student creativity grounded in the use of imagination or original ideas. Examples of creative topics include analyses of past events through the present lens or describing a person who has had a significant positive influence in a certain professional domain.

DIVERSITY AND INCLUSION

Prioritize content that supports diversity and inclusion and highlights depictions of race, ethnicity, language, national origin, religion, gender, age, disability, neurodiversity, sexual orientation, place of origin, beliefs, education, professional experience, communication style, and other cultural characteristics. For instance, topics in this category could address advantages of understanding and appreciating cultural differences and similarities. Avoid stereotyping when developing content related to diversity and inclusion.

RACIAL EQUALITY

Smarter Balanced is committed to being an anti-racist organization. Dismantling systemic and institutional racism is an essential issue that needs to be addressed in assessment content. Ongoing evaluation and reassessment of guidelines and testing materials serve to advance the collective understanding of racial injustice. Topics that promote racial equality and represent Black, Indigenous, Multiracial, People of Color (BIMPOC) and Black, Asian, and Minority Ethnicities (BAME) cultures are important to include. Topics that disrupt stereotypes or retell BIMPOC/BAME perspectives or narratives need to be prioritized. Contributions that are less common need to be included to avoid reinforcing the notion of exceptionalism of represented individuals. Fiction references that portray characters as individuals need to be selected. For instance, names representative of BIMPOC/BAME cultures need to be included.

EMPATHY

Topics that promote such socio-emotional skills as openness, awareness, and compassion toward diverse ideas, experiences, learning styles, and other personal, cultural, and universal characteristics should also be prioritized. Examples include supporting a friend who failed at achieving a set goal, supporting a person who has had a challenging day, or understanding what someone is experiencing when they are adjusting to a new culture.

AGENCY AND AUTONOMY

Prioritize topics that focus on modeling agency via activities that are meaningful and relevant to students, driven by their interests, and that are often self-initiated (with appropriate guidance from teachers). A possible example would be an item that depicts student agency giving students voice and choice in how they learn. Similarly, topics related to student autonomy that model situations in which students are wholly responsible for their decisions and implementation of subsequent actions are also desirable. For example, an item could describe how a student advocates for the accessibility resources they need in order to complete a test.

DIVERSE AUTHORSHIP

It is important to include content developed by authors of diverse backgrounds (ethnicity, race, gender, and others) that highlights those cultural perspectives and authentically represents those cultures. Examples include works authored by BIMPOC/BAME authors, authors of various gender identities, authors representing various nations, and authors with different abilities.

CULTURAL TERMINOLOGY

Smarter Balanced is committed to being an anti-racist organization while promoting accessibility through ongoing evaluation of test content and related materials to advance the collective understanding of such important dimensions of culture as race and ethnicity. Smarter Balanced uses appropriate capitalization for cultural groups and in so doing, acknowledges the tension of this stylistic choice when confronting the complex and challenging history of racialization in the United States of America. Smarter Balanced assessments strive to include a diversity of perspectives between and within groups. Whenever possible, test items should reflect the cultural identity of historical and present-day figures from diverse backgrounds in response to evolving sociocultural and sociohistorical conditions and avoid the implication that Whiteness is neutral or assumed.

When describing individuals from any group, use the label that the group prefers and allow room for self-identification whenever possible. Pay special attention to avoid derogatory culture-related terminology or labels. For example, avoid terms like “disadvantaged,” “oppressed,” “vulnerable” when describing individuals who identify as BIMPOC and BAME. Depending on the context, avoid using terms like “good,” “mainstream,” “average” when describing individuals with White cultural or ethnic identities. Avoid having people with Caucasian sounding names express more positive points of view, whereas names of those people who are from BIMPOC and BAME populations seemingly questioned for why they are right or not. The cultural identities indicated below offer specific guidance for test item creation, but do not represent an exhaustive list of groups that have been mislabeled throughout history.

BLACK AND AFRICAN AMERICAN PEOPLE

Use “Black” (capitalized) and “African American” and allow room for self-identification whenever possible. Test items and materials should take into consideration cultural experiences within the continent of Africa and across the African diaspora (e.g., Afro-Panamanian or Afro-Caribbean) or other national or tribal affiliations (e.g., Yoruba American or Nigerian American). Assessment content should be reflective of shared cultural history as well as racial/ethnic identity and experiences of Black people in the U.S. Avoid the pejorative use of “black” associated with evil, danger, or other negative aspects, as it may evoke stereotypes of behaviors or cultural characteristics.

ASIAN AMERICAN PEOPLE

When possible, use specific terms such as “Japanese American” or “Chinese American.” Terms such as “Pacific Island American,” “Native Hawaiian” (but use “Hawai‘i” for the name of the state) and “Asian/Pacific Island American” should be used as appropriate. Do not use the word “Oriental” to refer to people except in historical or literary material important for the measurement of state standards.

LATINO/LATINA/LATINX/LATINE AMERICAN PEOPLE

The terms “Latino American” (for a male), “Latina American” (for a female), and “Latinx” or “Latine” (for plural usage) are acceptable. The terms “Chicano,” “Chicana,” and “Chicanx” are also acceptable. When appropriate for the context, it is preferable to use specific group names such as “Cuban American,” “Dominican American,” or “Mexican American.”

NATIVE AMERICAN/ALASKA NATIVE PEOPLE

“Native American,” “Indigenous,” “First Nation,” “Alaska Native,” and “American Indian” are acceptable. When possible, use the specific tribal affiliated names for peoples such as “Pequot” or “Mohegan.” Some Native Americans prefer the words “nation” or “people” to the word “tribe.” If referring to a specific tribal nation or people, an approval is desirable from the tribal chairperson/leader. Additionally, this content would benefit from a review for accuracy by an expert from that tribal nation. There is great diversity among the 574 federally recognized tribes in the United States, including diversity in their languages, cultures, histories, and governments. Each tribe has a distinct and unique cultural heritage. It is important to understand that all land of North America is considered Indigenous homeland of Native American/Alaska Native/First Nation people. It is also important to ensure that content does not include stereotypes, misconceptions, and omissions about their history and identity in texts. Culturally responsive materials that honor and respect the identities of American Indian and Alaska Native people need to be prioritized.

LGBTQ+ PEOPLE

When considering topics related to lesbian, gay, bisexual, transgender, queer/questioning, and other identities (LGBTQ+), ensure that topics related to “outing” and bullying in association with these cultural groups are treated with care. Smarter Balanced is committed to ongoing evaluation and reassessment of guidelines and testing materials that serve to advance the collective understanding of the LGBTQ+ experience. Refer to people by orientation only when it is relevant to the construct being measured while using widely-accepted, inclusive terminology and allow room for self-identification wherever possible.

PEOPLE WITH DISABILITIES

Allow for optimal self-identification. Some individuals with disabilities prefer to put the person before the disability. For example, they use “a person who is blind” rather than “a blind person.” Other people with disabilities prefer to have their cultural affiliation to be stated first, e.g., “deaf people.” In general, avoid using adjectives as nouns for people with disabilities (e.g., “the blind” or “the deaf”) except in the names of organizations or in literary or historical material important for measurement of state standards. Avoid euphemisms such as “challenged.” Keep in mind that the term “neurodiversity” refers to the diversity of all

people, but it is often used in the context of autism. Use objective language rather than emotionally loaded terms (e.g., “uses a wheelchair” rather than “confined to a wheelchair”). Do not minimize disabilities by suggesting that they are not noticeable or important. Do not depict people with disabilities, including people with learning disabilities and people with developmental disabilities, as helpless victims. Do not state or imply that people with disabilities deserve to be pitied, feared, or ignored, or that they are more heroic, courageous, or special than people who do not have disabilities. Terms to be avoided include “dumb” for a person who is mute, “handicapped” for a person with a disability, and “retarded” for a person with a cognitive disability.

AGE

It is best to refer to people by specific ages or age ranges. Minimize the use of euphemisms such as “teens,” “young adults,” “seniors,” or “senior citizens.”

GENDER

Historical or literary material important for the measurement of state standards is acceptable even if it uses outdated terms when it comes to gender identity. However, the most common concepts of gender are based on the long-perpetuated notion that gender is a binary construct, and that it always aligns with a binary designation of sex (male/female). Yet, contemporary understandings of gender clarify that gender identity and expression occur along a broad spectrum that is not limited to two binary alternatives, such as woman/man or girl/boy (National Council for Teachers of English, 2018). Avoid male-dominated or binary language when talking about gender and use inclusive pronouns such as “they,” “them,” “theirs” instead. Use gender-neutral language, such as “spouses,” “parents,” and “partners.” Do not refer to women by physical attributes and to men by accomplishments as in “the successful lawyer and his beautiful wife.” Do not refer to males as “men” and females of similar ages as “girls.” Do not use the generic “he” or “man” to refer to all human beings. Avoid such gender-biased language as “mankind,” “man-made,” and “chairman.” Historical or literary material important for the measurement of state standards is acceptable even if it uses outdated terms and nonparallel language for women and men.

INTERSECTIONALITY

Forms of oppression and privilege, race, class, gender, sexuality, and other cultural dimensions intersect in people’s individual lives and in the surrounding environment. Intersectionality refers to the complexities of multiple marginalized identities, e.g., being Asian, female, disabled, and lesbian. It recognizes that each marginalized identity faces its specific challenges which, when combined, create unique and overlapping experiences of prejudice and discrimination. Avoid terminology that may stem from such experiences, e.g., “chief,” “rule of thumb,” “openly gay.”

REPRESENTATION OF DIVERSITY

There should be representations of different groups in the pool of items so tests built from the pool will, on average, be appropriately balanced. In items and stimuli that mention people, the following conditions are required in the pool of items and should be reflected in assignments to item writers:

- All genders should be considered and represented in assessment content.
- People of diverse backgrounds should be represented.
- People of different ages, abilities, and social classes should be depicted.
- People and contexts representing a variety of areas (urban, suburban, rural, etc.) should be included.
- A wide variety of life situations, living conditions, types of housing, types of families (including single-parent families), regions, and the like should be depicted.

A FINAL WORD

Neither this nor any other set of guidelines can cover all of the possible variations in content that will have to be evaluated for fairness in the Smarter Balanced assessments. Current events (e.g., pandemics, racial injustice, natural disasters, issues raised during political campaigns, terrorist attacks) can add new topics that may cause fairness problems. Issues that were neutral may become controversial. If the specific guidelines do not offer sufficient guidance in some particular situation, the best practice is to turn to the fundamental rules and ask:

- Do the items measure any irrelevant knowledge or skill? If so, will some group(s) be more greatly affected than others?
- Will any aspect of the test materials anger, offend, upset, or otherwise distract test takers? If so, will some group(s) be more greatly affected than others?
- Do the test materials treat all groups of people with respect? If not, will some group(s) be more greatly offended than others?

A potential fairness problem exists if some cultures are more greatly affected by particular test items than others. The next step is to determine whether or not the potential problem is a real one. Has difficulty been confused with fairness? Has a guideline been overgeneralized? Has a "treat carefully" guideline been interpreted as a "must avoid" guideline? Has a situation been contrived to make innocuous content seem unfair? Is the material important for valid measurement of state standards?

Finally, it is important to keep in mind that the intent of using the *Guidelines* is to remove unnecessary barriers to the success of diverse groups of test takers. Some potential barriers, such as complex language in an ELA stimulus based on historical documents, are acceptable and may be necessary to allow valid measurement of state standards.

REFERENCES

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

Bond, L. (1993). Comments on the O'Neill & McPeek paper. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277–280). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. (1989). Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszel method. *Applied Measurement in Education. 2,* 3, pp. 217–233.

Elliott, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment, 9*(1).

Holland, P. & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Johnstone, C., Altman, J., & Thurlow, M. (2006). A state guide to the development of universally designed assessments. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
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.
Ravitch, D. (2003). *The language police: How pressure groups restrict what students learn.* New York: Knopf.
Sireci, S. G. & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In *The History of Educational Measurement* (pp. 111–135). Routledge.
Smarter Balanced Assessment Consortium. (2021). *Smarter Balanced Assessment Consortium Usability, Accessibility, and Accommodations Guidelines.* Available from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf
Smarter Balanced Assessment Consortium. (2012). *Smarter Balanced Assessment Consortium accessibility guidelines for English language learners.* Olympia, WA: Author.
Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

OTHER USEFUL REFERENCES FOR FAIRNESS IN ASSESSMENT

ACT. (2011). *Fairness report for the ACT tests.* Iowa City, IA: Author.
American Institutes for Research. (n.d.). *Standards for language accessibility, bias, and sensitivity*. Washington, DC: Author.
American Psychological Association. (2001). *Publication manual of the American Psychological Association.* Washington, DC: Author.
Data Recognition Corporation. (2003). *Fairness in testing: Guidelines for training bias, fairness and sensitivity issues.* Maple Grove, MI: Author.
ETS. (2009). *ETS guidelines for fairness review of assessments.* Princeton, NJ: Author.
National Council for Teachers of English. (2018). Statement on gender and language. Available from https://ncte.org/app/uploads/2018/10/NCTE-Statement-on-Gender-and-Language.pdf?_ga=2.8457845.152354200.1597010915-1515076384.1582323058
Pearson. (2021). Pearson race & ethnicity: Diversity, equity, and inclusion guidelines (products). Available from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/news/DD02583PRIME-PearsonRaceEthnicityEquityDiversityandInclusionGuidelines002.pdf
Ramsey, P. (1993). Sensitivity review: The ETS experience as a case study. In P. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.
Zieky, M. J. (2006). Fairness review. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Mahwah, NJ: Lawrence Erlbaum Associates.

APPENDIX

EXAMPLES OF ACCEPTABLE AND UNACCEPTABLE TEST MATERIALS

The following examples are excerpts from test items and stimuli. Some of the excerpts illustrate various violations of the *Guidelines*. Others illustrate items that are acceptable in terms of fairness.

Math problems

The first set of examples consists of math problems or excerpts from math problems. Note the tension between adding realistic context to a math problem and avoiding linguistic complexity and irrelevant knowledge requirements.

1. The drawing below shows seismic bracing. Highlight the two triangles that are congruent with each other.
    a. **Unacceptable.** Few children are likely to be familiar with seismic bracing, and knowledge of seismic bracing is not related to the purpose of the question.

2. In the drawing below, highlight the two triangles that are congruent with each other.
    a. **Acceptable.** The reading load is reduced, and there is no unfamiliar context.

3. Shaquan helps assemble food packages for poor people at Christmas. Each box holds 6 cans in a row. There is room for 4 rows in a box. Write the expression that best describes the number of cans in one full box.
    a. **Unacceptable.** The first sentence adds to the reading load of a math question but adds no useful information. The references to “Christmas” and “poor people” are inappropriate and unnecessary.

4. Two people who were conversing at a street corner parted and moved away from the corner in straight lines that are perpendicular to each other. If one person walked at 3 miles per hour and the second person jogged at 4 miles per hour, how far apart would they be after one hour?
    a. **Unacceptable.** The linguistic load is high for a math question. The sentences are long, and the syntax is complicated. “Conversing” is a difficult synonym for “talking,” and the people’s actions before they started to move are not relevant in any case. (“Perpendicular” is acceptable as a valid mathematical term.)

5. Two people stood next to each other. They started walking in straight lines that are perpendicular to each other. One person walked at 3 miles per hour. The other person walked at 4 miles per hour. How far apart are they after one hour?
    a. **Acceptable.** Unnecessary information has been deleted. Two long sentences have been replaced by five shorter sentences. The conditional syntax (“If one person…”) has been replaced by brief statements of fact.

6. A modem can send X bits per second. Write the expression that shows how many seconds it would take to send Y bits.

a. **Unacceptable.** The mention of a "modem" and "bits" is irrelevant and is likely to be unfamiliar. A student might skip the item or waste time wondering what "bits" are or what a "modem" is.

7. Lee can walk X miles an hour. Write the expression that shows how many hours it would take Lee to walk Y miles.
   a. **Acceptable.** The context is familiar.

8. If one card is taken at random from a deck of playing cards, what is the probability that the card will be an ace?
   a. **Unacceptable.** The question assumes knowledge of the number of aces and the total number of cards in a deck of playing cards. It is acceptable to ask about probability, and it is acceptable to use playing cards in math problems. According to the guideline about gambling, however, it is not acceptable to assume that test takers have knowledge of the characteristics of a deck of playing cards.

9. There are 4 aces in a deck of 52 playing cards**.** If one card is taken at random from the deck, what is the probability that the card will be an ace?
   a. **Acceptable.** No knowledge of the characteristics of a deck of cards is required to answer the item.

10. When Ms. Luna pulled her car into the parking garage, the machine at the gate issued a ticket stamped with the time, 11:30 a.m. When she left the garage that afternoon, her ticket was stamped with the time she left, 12:15 p.m. What was the total length of time that Ms. Luna's car was in the parking garage?
   a. **Unacceptable.** The question is very wordy and uses an unfamiliar context for many children. In addition, "pulled her car" is an idiom that children may not know.

11. Sandip went to the library at 11:30 in the morning. He left at 12:15 that afternoon. How long did Sandip stay in the library?
   a. **Acceptable.** The reading load is reduced, and the context is familiar.

12. It takes Sarah an average of 30 minutes to clean her bedroom. She cleans her bedroom once a week. How many hours would Sarah spend cleaning her bedroom in one year? Acceptability depends on the mix of items in the test.
   a. **Unacceptable** if many questions in the test had girls cleaning rooms or doing what was traditionally considered "woman's work," because the test would reinforce a stereotype and be unfair.
   b. **Acceptable** if combined with questions showing women doing nontraditional work. Not all children have their own bedrooms, but the concept that some children have individual bedrooms should be neither strange nor upsetting. Whether or not the required knowledge of the number of minutes in an hour and the number of weeks in a year is fair depends on the grade level of the test takers.

13. According to the graph, the number of unemployed workers was highest in which year?
   a. **Acceptable.** The mere mention of unemployed workers is acceptable.

14. Marisa hit the bull's-eye with her rifle 7 times out of 9 shots. What percent of the time did Marisa hit the bull's-eye?

a. **Unacceptable.** Students who are not familiar with the phrase "bull's-eye" in the context of a target will have a rather gruesome mental picture of Marisa's shooting if taken literally. The use of guns tends to be controversial in any case.

15. The data tables below show how long a driver will be impaired based on the consumption of 1, 2, or 3 ounces of alcohol within one hour. Use the data to predict the amount of time a driver will be impaired after consuming 4 or 5 ounces of alcohol in one hour. Explain your reasoning for obtaining the predicted values.

a. **Unacceptable**. A brief item concerning alcohol might be acceptable in the higher grades in the context of showing impairment, but basing an entire performance item on the topic is excessive. Also, showing consumption of more than 1 or 2 drinks per hour models inappropriate or even dangerous behavior.

16. A pizza is cut into 8 slices. If 5 students eat one slice each, how many slices will be left?

a. **Acceptable.** One slice of pizza is not excessive consumption of food.

Excerpts from stimuli for English language arts

The next set of examples consists of brief excerpts from ELA stimuli.

1. Wagner used the orchestra to achieve certain effects in much the same way that other composers of operas used the singers.

a. **Acceptable** if the knowledge needed to answer the questions was included in the passage. The mere mention of opera or a composer does not make the excerpt unfair.

b. **Unacceptable** if understanding the passage required knowledge of opera and how composers "used" the orchestra or "used" singers.

2. The African Americans living in Middletown tended to be part of households consisting of extended families living together.

a. **Acceptable.** The statement of fact about a particular group of African American people is acceptable and does not reinforce a stereotype.

3. Cyanide is one of the fastest-acting poisons known to science.

a. **Unacceptable.** The excerpt violates the guideline about avoiding dangerous actions and substances. Parents are likely to oppose including information about lethal substances in the test.

4. The AIDS epidemic, which has devastated some countries in sub-Saharan Africa, has affected children as well as adults, leaving many children not only orphaned and uncared for, but also malnourished, diseased, and close to death.

a. **Unacceptable.** Excessive detail about the suffering of children makes the excerpt unacceptable.

5. Harlow was best known for the experiment in which he separated infant monkeys from their mothers shortly after the infants were born.

a. **Unacceptable.** The excerpt violates the guideline that prohibits inclusion of painful or harmful experimentation. The excerpt would be acceptable in a psychology test, however.

6. I love to make videos! I use the camera in my phone to capture my friends having a good time with their dates at parties and at school dances.
    a. **Unacceptable.** Owning a cell phone with video capabilities is currently a luxury beyond the reach of many test takers. The references to "dates" and "dances" are not in compliance with the guideline concerning social dancing.

7. An ancestor of the modern horse the size of a dog gave rise to progressively larger species.
    a. **Acceptable.** The passage concerns the evolution of horses, which is in compliance with the guideline that identifies the evolution of human beings as the aspect of evolution to avoid.

8. The Japanese immigrants enrolled in Ms. Kubota's class worked very hard.
    a. **Acceptable.** The reference is to a particular group of Japanese immigrants, so it does not stereotype all Japanese immigrants.

9. The amount of caffeine in a cup of coffee can still affect the human body more than three hours after it has been ingested.
    a. **Acceptable.** The mention of caffeine appears to be in an objective discussion of the effects of drinking coffee and would be in compliance with the guideline on harmful substances, if the passage did not encourage the drinking of coffee.

10. People who drive gas-guzzling SUVs contribute to global warming.
    a. **Unacceptable.** This excerpt is a clear violation of the guideline against advocating for one side in a controversial situation.

11. In the 17th century, many convicted criminals were hanged, but some were slowly crushed to death.
    a. **Unacceptable.** Death by slow crushing is clearly out of compliance with the guidelines about death and suffering.

12. A large influx of immigrants will destroy the equilibrium of a neighborhood.
    a. **Unacceptable.** The negative view of immigrants in the excerpt makes it out of compliance with the guideline forbidding offensive stereotypes of any group. The verb "destroy" is particularly harsh in that context.

13. There has been an increase in the number of people who identify themselves as American Indians.
    a. **Acceptable.** Either "American Indian" or "Native American" is appropriate. The fact that more people than before identify themselves as American Indians is not a fairness problem.

14. Surprisingly, a girl won the math contest.
    a. **Unacceptable.** By expressing surprise that a girl won the math contest, the excerpt reinforces the stereotype that girls have less quantitative ability than boys.

15. The soldiers and their wives attended the ceremony.
    a. **Unacceptable.** Unless the reference is to a previously specified group of all male soldiers, refer to "the soldiers and their spouses" to avoid the implication that only males are soldiers.

16. ...that all men are created equal, that they are endowed by their Creator with certain unalienable Rights…

    a. **Acceptable.** In spite of the use of "men" to refer to all people and in spite of the reference to God, the excerpt is acceptable because it is from an important historical document of the type required by state standards.

17. Bridges with steel frames are more likely to survive an earthquake than are stone bridges.

    a. **Acceptable.** The mention of a natural disaster is acceptable. There is no focus on death and destruction.

18. Lee's father and Juan's father are both policemen.

    a. **Unacceptable.** Even though both officers are male, "police officers" is preferred to "policemen" to avoid the impression that only men are police officers.

19. The ancient Romans played handball and engaged in other sports while nude in the public baths.

    a. **Unacceptable.** Though unintended, "engaged in other sports while nude" could be taken as sexual innuendo.

20. Many of the people in the United States who speak Spanish come from Mexico.

    a. **Acceptable.** The excerpt is a statement of fact and is not a violation of any guideline.

21. He be at work…

    a. **Unacceptable.** The use of dialect is stereotyped language and is in violation of the *Guidelines* unless it is important to measure a state standard in literary or historical material.

22. The men's room is on the right; the girls' room is on the left.

    a. **Unacceptable.** Parallel language would call for "women" to match "men" or "boy" to match "girl."

23. Some Native Americans claim to be members of the Algonquian tribe, but according to anthropologists, "Algonquian" is a general term applied to many Native American peoples who speak related languages, not the name of any particular tribe.

    a. **Unacceptable.** There is a problem in that the academic definition of "Algonquian" is taken as correct, but the usage of Native Americans about themselves is taken as incorrect. The excerpt is out of compliance with the guideline to call people what they prefer to be called.

24. Frederick Douglass, the great African American abolitionist, was said to be born on Valentine's Day.

    a. **Acceptable.** The excerpt requires no knowledge of how or why Valentine's Day is celebrated, nor any agreement that it should be celebrated.

25. Edward Said and Daniel Barenboim cofounded a children's orchestra.

    a. **Unacceptable.** Though there is nothing overtly problematic about the excerpt, Edward Said was famous as a Palestinian activist and remains a highly controversial figure. He is viewed very positively by some groups and very negatively by other groups. Reviewers who are not familiar with the people depicted in test materials should check reference sources to avoid the inadvertent inclusion of controversial figures.

26. The President issued a Proclamation freeing enslaved Africans, stating, “that on the 1st day of January, A.D. 1863, all persons held as slaves within any State or designated part of a State the people whereof shall then be in rebellion against the United States shall be then, thenceforward, and forever free.”

a. **Acceptable.** Mention of slavery is acceptable, and state standards call for the inclusion of documents important in American history, such as the Emancipation Proclamation.

English language arts items

The next set of examples consists of items or parts of ELA items.

1. According to the passage, how long ago did Homo sapiens evolve into a distinct species?

a. **Unacceptable.** To answer the question about when human beings evolved implies that human beings evolved from other species. That implication is not in compliance with the guideline regarding evolution.

2. The character delivering the monologue attributes the arrogance of the French to which of the following?

a. **Unacceptable.** Describing all of the people in a nation as “arrogant” is a clear case of offensive stereotyping. The question is not in compliance with the guideline concerning stereotypes.

3. Describe the changes within the ecosystem portrayed in the video, including the impact of man’s activities on weather patterns, and possible solutions to correct ecological problems.

a. **Unacceptable.** The question uses “man” to refer to all people, which is not in compliance with the guideline on appropriate terminology for men and women. The influence of people on climate change is controversial and out of compliance with the guideline on the avoidance of advocacy.

4. The author compares the artist’s use of color to which of the following?

a. **Acceptable** if direct experience of color is not required to understand the passage and answer the items.

b. **Unacceptable** if direct experience of color is required. The material would be unfair for students who are blind.

5. Our society stereotypes old people as weak, uninformed, forgetful, and foolish. Discuss the extent to which you agree or disagree with this stereotype.

a. **Unacceptable.** The question blatantly reinforces stereotypes of a group and invites test takers to agree with the offensive stereotypes.

6. Isaiah wrote, “Woe unto them that are wise in their own eyes.” Describe the meaning of that quotation and give two examples of people who are “wise in their own eyes” from your reading or from your personal experience. Explain your choices.

a. **Unacceptable.** The excerpt violates guidelines about the avoidance of religious material, even though the students are not asked to write directly about religion.

7. In the play, Luz was restricted to a wheelchair for which of the following reasons?

a. **Unacceptable.** The phrase "was restricted to a wheelchair" should be replaced with more objective terminology such as "began using a wheelchair."

8. According to the newspaper article, Robert died how many years after his brother John?
   a. **Acceptable.** According to the *Guidelines*, it is acceptable to mention death as long as gruesome details are not depicted.

9. It can be inferred from the passage that the spinnaker is most effective during a race when the wind is in which position relative to the boat?
   a. **Unacceptable.** Using sailboats for racing is out of compliance with the prohibition against luxuries. Also, unless "spinnaker" and its use are explained clearly in the passage, the item would depend on irrelevant specialized knowledge.

10. Based on information in the documentary, which of the following people is most likely to carry the sickle cell trait but show no symptoms of the sickle cell disease?
   a. **Unacceptable.** Diseases that affect particular groups of people are likely to be problematic in terms of fairness. This topic is best avoided.

11. The lecturer stated that among spiders found in many houses in the United States, the bite of which of the following is most likely to cause painful, deep wounds?
   a. **Unacceptable.** The focus on "painful, deep wounds" from spiders "found in many houses" makes the item out of compliance with the guideline regarding animals that are frightening to children.

12. The video excerpt of Baryshnikov dancing in *The Nutcracker* best illustrates which of the following aspects of his work described in the magazine article?
   a. **Acceptable** if all of the information needed to respond to the item is included in the video excerpt and the magazine article.
   b. **Unacceptable** if knowledge of ballet is required to answer the item. Only social dancing of couples is prohibited by the *Guidelines*.

13. Read the excerpt from the diary of a ship captain engaged in transporting slaves and watch the video dealing with the history of slavery in the United States. Imagine that you are a newly captured slave. Describe your experiences on land and on the sea during your journey from Africa to the United States. Use information from both the diary and the video in your description.
   a. **Unacceptable.** Mention of slavery as a topic is acceptable, but forcing test takers to imagine that they personally experienced the transatlantic journey, during which many captives are known to have suffered and died, will be upsetting to some students.