Fairness and Sensitivity Guidelines

2025

Note on the current edition

This revision of the *Bias and Sensitivity Guidelines* marks both a culmination of many years of experience and a departure from the past. The new title is the first and most apparent change in this document: *Fairness and Sensitivity Guidelines*. This is an intentional shift from avoiding bias to focusing on promoting fairness. This document uses an asset-based framework to connect to the core motivation for these guidelines: to promote an awareness of how test content and design can aim toward a more fair and equitable assessment.

As with this document, the notion of fairness in assessment itself is continuously evolving (Russell, 2024; Sireci & Randall, 2021). These *Fairness and Sensitivity Guidelines* (referred to as the *Fairness Guidelines* for the remainder of this document) are reviewed by an external panel of experts every two years and revised to reflect the current needs and understanding of fairness and equity for all students. The changes in the current edition reflect the review committee's input and the current direction of Smarter Balanced.

The *Fairness Guidelines* were revised for flow and sequence. The purpose of this revision was to make the document more accessible and usable for readers. This version seeks to maintain the essence of the previous guidelines while streamlining and creating an organizational flow to support comprehension and application of these ideas. The guidelines end with a short checklist that can be used during the test content development and review process as a reminder of the considerations addressed in this document.

TABLE OF CONTENTS

INTRODUCTION

The *Fairness and Sensitivity Guidelines* support the development of Smarter Balanced assessments that are fair and accessible for all test takers. Fairness in assessment means ensuring that test materials are free, as much as possible, from unnecessary barriers that could impact the success of diverse groups of students. This means all assessments must include test content (stimuli, items, and tasks) that:

- Focuses on measuring only relevant knowledge and skills.
- Is grounded in respectful representation and treatment of diverse populations and experiences.
- Does not advantage or disadvantage students.
- Does not upset or distract test takers.

This document describes in detail how to follow these guidelines for the Smarter Balanced assessments of the state standards in English language arts/literacy (ELA) and mathematics. The *Fairness Guidelines* have been agreed upon by Consortium advisories after undergoing reviews with input from diverse experts and the latest findings and trends in the field.

The *Fairness Guidelines* must be considered in all phases of test design, development, and use, including item writing and reviewing, adding accessibility resources, reviewing post-field test data, and selecting operational items. These guidelines provide the foundation for item writers and reviewers to address multiple viewpoints when evaluating items on Smarter Balanced assessments.

Use of the *Fairness Guidelines* helps ensure that the Smarter Balanced assessments comply with Standard 3.2 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) (referred to as the *Standards* for the rest of this document):

> *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.* (*Standards*, p. 64).

All test materials to be included in Smarter Balanced assessments must adhere to the *Fairness Guidelines*.

WHAT IS FAIRNESS?

Defining Fairness

"Fairness" can be a challenging word to define because it is used in many ways. There is not one single definition of fairness, and no one test will ever be completely fair to every test taker at every moment. The goal is to strive for fairness to the maximum extent possible, focusing on access and opportunity for all student populations. When thinking about fairness in assessment, we refer to the foundational definition offered by the *Standards:*

> *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.* (*Standards*, p. 50).

What Fairness is Not

A common misperception is that if a test item is difficult, or more challenging for some students than others, it is inherently unfair. However, difficulty alone does not determine fairness. An item is considered fair if the sources of difficulty are valid and relevant to the assessment's purpose. For instance, state standards often require students to read primary source documents from United States history, which can be challenging. Valid test items will reflect this difficulty appropriately. Fairness is partly determined by the purpose of the item or task. Even if some of the items are more difficult for certain groups of students, this does not automatically make them unfair.

For example, you might wonder if English language reading comprehension items are fair for all students, including those who are multilingual learners. It is likely that students will vary their performance on test items depending on their varying levels of English language proficiency. Differences in these scores, in this case, are fair because they represent the purpose or construct of the item—to test English reading comprehension skills. In other words, score differences between groups based on actual differences in comprehension of English match the intention of the test and are, therefore, fair.

This rationale about fairness extends to examining differences in test scores between student groups. Fairness does not require that all groups have the same average scores. Fairness requires any existing differences in scores to be valid or representing the purpose or construct of the item.

Fairness can also vary based on the content of the test item, not just the skills it intends to measure. 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 upsetting and, thus, were unable to use their attention and energy to focus on the content of the test. These concepts and examples are explored in depth in later sections of this document.

Fairness and Validity

Fairness and validity are related ideas that, together, are building blocks for a high-quality assessment. While fairness refers to the test materials, validity focuses on what meaning is attributed to the test scores.

Test validity is whether the inferences made based on test scores are appropriate and backed by evidence (Messick, 1989). More simply, validity is 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. Validity of the resultant test scores measuring these standards rely on both alignment and fairness of test content.

Integrating fairness into the *foundations* of assessment development is an essential step to ensuring test validity. It is a proactive, and not a reactive approach that considers fair test design as part of seeking justice for marginalized communities (Randall, et al., 2022). The steps presented in these *Fairness Guidelines* play a vital role in ensuring the authentic and appropriate representation of varied student experiences in assessment content, as well as supporting valid interpretations of what all students know and can do.

Fairness and Assessment Design

Validity starts from the blueprints for test item development. The Smarter Balanced assessments are developed using the principles of Evidence-Centered Design (ECD) (Mislevy, Steinberg, & Almond, 1999). Three basic elements of ECD are: 1) clearly 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. ECD provides a chain of evidence-based reasoning linking test performance to the claims to be made about test takers.

Fair assessments are both an essential input and outcome of ECD. If test items are not fair, then the evidence they provide is not an accurate representation of the skills and knowledge of all students. Under these circumstances, the claims cannot be equally well-supported for all test takers. Therefore, appropriate use of the *Fairness 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.

In addition to ECD, Principles of Universal Design underlie the design of Smarter Balanced assessments (Johnstone, Altman, & Thurlow, 2006). These principles promote the following:

- 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, comprehensibility, and legibility

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. These principles help reduce visual, auditory, cognitive, communicative, physical, and other barriers to valid measurement.

Accessibility Supports

The Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* (UAAG), 2024, and the Smarter Balanced *Accessibility Guidelines for English Language Learners* (2012) provide information about how to ensure that the Smarter Balanced assessments are as accessible as possible for all, including English learners, students with disabilities, English learners with disabilities, and other general education students.

Test content developers and reviewers need to be aware of the different supports that students may use when taking Smarter Balanced assessments.

Accessibility resources can be embedded or non-embedded and fall under three main categories:

- Universal tools
- Designated supports
- Accommodations

Accessibility needs and preferences are unique and can be customized for participating general education students, not only those who may have an English learner or disability status. Embedded and non-embedded universal tools allow for this flexibility.

However, even with Universally Designed assessments, some designated supports and accommodations are still necessary. For example:

- Math content and ELA auditory stimulus require American Sign Language (ASL) translations.
- Braille transcriptions of images require screen readers and embossing.

Language Complexity

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 on test content that is not directly related to measuring language skills, such as mathematics. The language in mathematics items should not be a barrier to a correct answer for students who could otherwise do the required mathematics. For mathematics assessments, non-mathematical language should be clear, simple, and targeted no higher than at the tested grade level.

The focus of ELA assessments is on language use, although unnecessary complexities must be avoided in this content area too. Valid assessment of ELA state standards requires language targeted at the tested grade level. In general, the clearest language consistent with validity (see above) should be used when the language itself is not being tested. Smarter Balanced Content and Item Specifications are designed to promote these principles of Universal Design and accessibility.

HOW IS FAIRNESS EVALUATED?

Evaluating fairness is a complex, multi-step process. No one person can judge for fairness, and no one metric provides the comprehensive review for fairness that is required to achieve validity. Instead, a mixed-methods approach, with both qualitative and quantitative data, is required in evaluating test content for fairness. 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.

Fairness evaluations begin with trained reviewers with diverse experiences and perspectives. Issues that may affect fairness are often too subtle to be captured by any statistic. The content and examples in these *Fairness Guidelines* form the core of that review process.

Test developers may, however, miss potential fairness issues that sophisticated statistical analyses later find (Bond, 1993). 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 (*Standards,* pp. 65-66).

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 "matched." Significant differences in test item difficulties or matched people in different groups result in higher DIF statistical values.

> 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.

DIF analyses are limited when it comes to identifying biases and other research methods need to be explored. 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 differ on the relevant knowledge that an item is supposed to measure. Due to lack of opportunity to learn or other factors, 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 alone, however, is not proof of bias. No test is perfect. Therefore, matching test takers based on test scores cannot be perfect. A fair item may show DIF merely because the test scores have not matched people well regarding 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 of a test item

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 or task type.

The remainder of this guide supports the qualitative review of test content for fairness.

CONTENT TO PRIORITIZE

Eligible Content

Fair assessments start with the test content itself. This is the first step, and an essential one, to building an assessment that has minimal barriers to fairness. Eligible content on Smarter Balanced assessments includes content that is:

- required by state standards,
- consistent with the Consortium's item specifications guidelines, and
- reviewed through item review procedures.

State or territory level laws and policies

State laws and state policies in one or more of the member states or territories (state-specific content) may also affect the assessment content that they present to their students. State-specific content may be administered during the same test event as a Smarter Balanced assessment if it conforms to the Consortium's policies regarding adding 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.

Test content needs to focus on students' ability to demonstrate independence, respond to various audiences and tasks, analyze evidence, understand other perspectives, as well as other elements included in state standards. The following are content topics to prioritize in assessment development:

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 in Authorship

It is important to include content developed by authors of diverse backgrounds that highlights different cultural perspectives and authentically represents those cultures.

Prosocial Attitudes

Topics that promote skills such as openness, awareness, and compassion toward others' 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). An example would be an item that depicts agency by giving students voice and choice in how they learn. Similarly, topics related to student autonomy that model situations in which students are responsible for their decisions and implementation of subsequent actions are also desirable.

Environmental Awareness and Sustainability

Highlight the significance of ecological responsibility, and the importance of understanding human impact on the environment. This would encompass such topics as conservation, renewable resources, and the role of technology in sustainability. For instance, an item could address the benefits of community-based recycling programs.

FAIRNESS CONSIDERATIONS

One fairness concern is that students differ in exposure to information through their life experiences outside of school. These variations can be due to geographic and demographic differences in where students live and how they spend their time outside of school. This variation in exposure to information becomes relevant when choosing discretionary assessment materials. State standards do not include all the content areas from which topics and contexts must be drawn. Stimuli (i.e., reading passages) for ELA items must be about a specific topic to provide a prompt for students to read and respond to associated items. Mathematics problems are often placed in real-world contexts to measure mathematical reasoning.

Which topics and contexts are fair to include in the Smarter Balanced assessments?

Even though curricula differ across and within states, 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. The key here is to stay aligned with grade-level standards and concepts that students learn about in school.

Here are some examples of concepts that a student might reasonably learn about in school, even if they may not have direct experience with them in their personal lives:

- Weather features such as snow, rain, sunshine, clouds, etc.
- Geographical features such as oceans, mountains, deserts, etc.
- Living accommodations such as a house, apartment, farm, etc.
- Activities such as cooking, running, reading, biking, swimming, etc.

Not all assessment content needs to be a repeat of what students have already had exposure to in school. More specifically, a major purpose of reading is to learn about new things. Therefore, it is acceptable to include material that may be unfamiliar to students *if the information necessary to answer the items is included in the tested material*. Selecting a reading passage that is likely new to students is therefore allowed on Smarter Balanced assessments, if the test items do not require outside knowledge of the topic beyond the stimulus itself.

It is helpful to employ a *window-mirror* approach to understanding how familiar students need to be with assessment content. The stimulus or item content serves as a *window* to the topic for the student. The task of the student is to *mirror* that content back and respond to the test item meaningfully.

Content Accessibility

A similar consideration about familiarity with assessment content arises for students with disabilities. For example, there may be questions about 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 fair to include material that may be unfamiliar to some students based on their life experiences, if 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*.

Some content is unfair because it conveys ableist assumptions about disability. 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.

Diverse Representation

Representations of different groups in the pool of items ensures that tests built from the item 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 must be reflected in assignments to item writers:

- 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 and regions (urban, suburban, rural, etc.) should be included.
- A wide variety of life situations, living conditions, types of housing, and family structures (including single-parent families) should be depicted.

Irrelevant Content

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.

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, a test item about 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.

It is necessary to avoid unfair barriers to success that are 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 (e.g., how to subtract) may not have the irrelevant knowledge of hockey needed to answer the item (e.g., a math item that states three members of the team were absent and asks how many were at the game).

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 fair, if understanding the explanation or making inferences from the explanation is the knowledge being assessed.

The following categories are common sources of specialized irrelevant knowledge in some large-scale assessment content (see Table 1 on the following page). 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.

Table 1: Sources of Irrelevant Content Knowledge

Source	Description
Regionalisms	Avoid 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	Avoid requiring knowledge about any one religion as well as content that privileges a particular religion or system of beliefs over others. For example, to say that something is "as colorful as an Easter egg" may be an unfamiliar comparison for some students.
Occupational & Technical Information*	Avoid 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."
Slang	Avoid highly contextual slang terminology that may be unfamiliar 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*	Avoid the overuse of academic terms in construct-irrelevant parts of an item, as this may interfere with assessment validity. For example, instructions or directions that use "register your answer" instead of "enter your answer."
Figures of Speech	Avoid 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, understanding 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 & Technology	Avoid 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. These terms can 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).

* The point at which words become overly specialized or overly academic 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.

Stressful or Upsetting Content

Barriers stemming from stressful content 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, 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 because of stressful events that may have interfered with their sense of security. For example, content related to pandemics or natural disasters must be addressed in a way that ensures those topics do not trigger or traumatize students.

Even if student performance is not directly affected, the presence of micro-aggressive, 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.

Certain topics are 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 or intentionally and thoughtfully address critical issues. 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. Test content that may cause strong negative emotions such as anger, disgust, fear, hatred, or sadness should be avoided.

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 as current and emerging events may add topics. Current and emerging events may add topics that are so problematic that they should be excluded from the assessments. Topics to avoid 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 innuendo
- Suicide
- Torture

Political Topics

There are also several topics that need to be approached thoughtfully, particularly those topics that pertain to sensitive or debate-worthy issues. Although such topics can be included for historical purposes or with appropriate contextualization , 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 actions
- 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 approached thoughtfully to minimize potential fairness issues. Guidelines that forbid a topic are easy to apply. Guidelines that require approaching a topic thoughtfully are more difficult to apply because different people will have different opinions about what is fair. 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 be fair and valid for test takers, and they must appear to be fair and valid according to the various constituencies within the Consortium. It is counterproductive to use test materials that various groups within the Consortium will consider inappropriate for their students.

Table 2: Topics to Avoid or Approach Thoughtfully

Topic	Description
Accidents and Natural Disasters	Avoid focusing on suffering, destruction, loss of life, loss of property, or graphic, gruesome details of damage caused by accidents and natural disasters that may upset or frighten students.
Advocacy or Lobbying	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	Avoid depictions of people using alcohol, tobacco, and illegal drugs. Do not depict the use of these substances as pleasurable, alluring, or as signs of sophistication and maturity. Warnings against the use of these substances may be fair for students in middle or high school.
Animals Frightening to Children	Younger students are more likely to be upset by certain dangerous animals than older students. Avoid depictions of spiders and poisonous snakes because it can cause problems for some children. Objective depictions of a food chain or non-threatening descriptions of animals are fair, 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 fair. Avoid a depiction of a wolf ripping the entrails from a fawn or attacking a child.
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 or thereafter.
Current Events & Contemporary Persons	Content depicting events related to or people involved in negative contexts, such as criminal activity, violations of human rights, etc., should be approached thoughtfully. Special attention should be paid to the extent to which controversial content is detailed and whether the content is generally appropriate.

Topic	Description
Dancing	Avoid couples dancing in a social setting, which is the form of dance that will most likely draw criticism.
Dangerous Activities	Avoid modeling behaviors that are inherently dangerous, or making dangerous behaviors appear 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 fair if depicted as being done properly. It is not fair to describe dangerous substances or devices such as weapons, poisons, or explosives in ways that make them appear attractive or safe.
Death & 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 fair to mention death (e.g., Rosa Parks died in 2005), but it is not fair to depict gruesome details.
Evictions & Unhoused Conditions	Discussion of evictions and unhoused conditions may be particularly upsetting to students who have lived experience with evictions and being unhoused or fear experiencing them in the future. These topics must be treated factually and thoughtfully. Avoid the emotional discussion of these topics, including aspects that focus on anguish and distress.
Evolution	Approach the topic of evolution of human beings or similarity of human beings to other primates thoughtfully because it is highly controversial for some people. Evolution within a species (such as evolution of bacteria to withstand antibiotics) is much less problematic and could be allowed if approached thoughtfully. Fossils and the age of Earth are fair if not linked to the evolution of human beings. (In tests intended to measure knowledge of science, any aspect of evolution required to measure this construct is fair.)
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.

Topic	Description
Food & Dietary Choices	Dietary choices are deeply personal and can be influenced by a multitude of factors, including health conditions, religious beliefs, ethical considerations, cultural practices, and personal preferences. Mocking or marginalizing someone's dietary choices can be perceived as disrespectful or ignorant. For example, it can be noted that someone "follows a vegan diet" rather than referring to their diet as "picky."
Gambling	Instruments used for gambling, such as playing cards and dice, may be used as required in mathematics 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 harmful behavior for students so as not to upset anyone who may have experienced this behavior. Examples of harmful, criminal, or inappropriate behaviors include bullying, cheating, truancy, joining gangs, fighting, lying, and stealing. It is particularly important to avoid making this behavior appear attractive, fun, glamorous, sophisticated, or something to be emulated. Instead, consider prosocial examples such as making amends for wrongdoing, accepting responsibility for one's actions, making changes to improve harmful behaviors, being reliable, and truthful.
Historical Figures, Events, and Places	Approach thoughtfully narratives related to historical figures that explicitly or implicitly point to those figures' involvement in harmful contexts, such as criminal activity. 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.
Holidays & Birthdays	Mentioning holidays and birthdays is fair if all 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	If the topic of immigration is included to measure a state standard, it should be addressed factually, objectively, and with sensitivity. Some test takers or their families may have personal experiences related to immigration that could include difficult or complex circumstances. To ensure inclusivity and respect, item developers are encouraged to present scenarios that reflect a broad range of cultural perspectives and experiences.
Luxuries	Avoid discussion of luxuries or the accumulation of wealth rooted in the exploitation of labor and/or natural resources, unless needed to measure state standards in literary or historical materials. Avoid depicting expenditures that would distract from the state standard being assessed.

Topic	Description
Medicines & Wellness	Approach medical treatments for serious illnesses thoughtfully, as this topic may be upsetting to some students and/or groups who are opposed to medical treatment. Do not model the use of drugs, including (but not limited to) prescription drugs and/or dietary supplements. Instead, focus on behaviors that promote wellness while paying special attention to issues associated with ableism.
Mental Health	Mental health conditions can shape personal experiences and interactions. Mental health is a sensitive topic and needs to be approached thoughtfully. 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,” and “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 depression or anxiety. Consider modeling mindfulness, self-care, and empathy instead.
Family Structures	Family structures and dynamics have evolved significantly over the years, and it is essential to recognize that there is no ‘one-size-fits-all’ definition of family. Stigmatizing certain parental statuses (e.g., portraying single parenting as less valued than a dual-parent household) can perpetuate harmful stereotypes and trigger negative feelings in students.
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 & Attributes	Avoid upsetting students by depicting their heights, weights, or other physical attributes with negative connotations. When developing test items, it is crucial to avoid reinforcing harmful beauty standards or body-shaming, which can perpetuate negative feelings. A wide range of body types must be represented in any written, oral, or visual material, but take care to avoid stereotypes and negative depictions of body shapes and other physical or psychological characteristics.
Pregnancy	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, or speaking with friends in a park).

Topic	Description
Religion	Religion is a source of information that is not common or accessible to all students. In Smarter Balanced assessments, religion is a topic that needs to be approached thoughtfully. Some people will see even an objective description of a religion as proselytizing. However, it is fair to mention religion. For example, noting that Buddhism is one of the main religions in Singapore is fair. Going into detail about the practices of adherents of Buddhism is not fair. 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 Illness	Serious illnesses include mental as well as physical illnesses. Illnesses that primarily affect certain groups, such as some genetic diseases, may be particularly problematic. Mentioning serious illnesses may be fair, 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 approached thoughtfully when it comes to depictions of negative social media effects on youth mental health and behaviors, such as cyberbullying, harassment online, and social exclusion. Special attention must 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, War, Violence & 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.

The challenge of overextension

There are many factors to be considered when ensuring the integrity and fairness of content selected and represented on Smarter Balanced assessments. Therefore, consider the *Fairness Guidelines* as a helpful tool to review for fairness. We want to avoid situations where we make up extreme situations in which a relatively innocuous topic is judged to be unfair. 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 set of circumstances is not necessarily unfair.

Avoiding Stereotypes

Materials in Smarter Balanced assessments must avoid stereotyping. For example, it is fair to depict gender-conforming behavior (e.g., a woman caring for children), but gender-conforming behaviors must be balanced by depictions of gender-nonconforming 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 behaviors whenever they produce items showing gender-conforming behaviors that could be considered stereotypes.

Misuse of language is also a concern. Special attention should be paid to cultural misnomers, such as treating Africa as a country or using *American* where *U.S. American* would be more appropriate. 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" to achieve greater diversity in the item pool. Finally, it is important to prioritize ethno-relative (culture-general) content and either avoid or sufficiently describe ethnocentric (culture-specific) content. For instance, if the item is based on the rules of soccer, those rules should be clearly outlined in the item.

(See Table 3 on the next page.)

Table 3: Examples of Stereotypes in Item Language and Content

Stereotype	Description
Language	Some stereotyped language may be fair 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 fair. 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 unfair (e.g., use "sales representative" instead of "salesman," "firefighter" instead of "fireman," "mail carrier" instead of "mailman").
Images	Do not use images that perpetuate stereotypes of people by sex, race, ethnicity, class occupation, etc. For example, avoid images that 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 must be shown across images.
Social and Occupational Roles	Individuals from different backgrounds should be represented in different occupational roles with varying levels of power. There must be a mix of sex, race, 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. Do not depict all male doctors with all female nurses. Diversity of roles must be shown across items if it is impossible to show it in a single item.
Behaviors and Characteristics	Reflect the diversity of the human experience. Test items and materials must not present the characteristics, traits, or lived experience of any group as a monolith. 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. Do not stereotype or overgeneralize specific actions, behaviors, or beliefs to a particular subgroup of individuals based on identities.

CULTURALLY APPROPRIATE TERMINOLOGY

Smarter Balanced is committed to promoting accessibility through ongoing evaluation of test content and related materials. Smarter Balanced assessments strive to include a variety of perspectives between and within groups.

When describing individuals from any group, use appropriate capitalization for cultural groups; use the label that the group prefers, and allow room for self-identification whenever possible. Special attention must be paid to avoiding derogatory culture-related terminology or labels. For example, avoid terms like “disadvantaged,” “oppressed,” or “vulnerable” when describing individuals who identify as Black, Indigenous, Multiracial, and People of Color (BIMPOC), and Black, Asian, and Minority Ethnic (BAME). Depending on the context, avoid using terms such as “good,” “mainstream,” or “average” when describing individuals with White cultural or ethnic identities. The cultural identities indicated in table 4 offer specific guidance for test-item creation, but do not represent an exhaustive list of groups.

Table 4: Culturally Appropriate Terminology

Terminology	Description
Black and African American People	Use "Black" (capitalized) and "African American" and allow room for self-identification whenever possible. Test items and materials must consider the range of 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 can 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" can 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 & Chicano/Chicana American People	The terms "Latino American" (for a male), "Latina American" (for a female) are fair. The terms "Chicano" and "Chicana," are also fair. 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," and "Alaska Native" are fair uses. 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." 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 ensure that content does not include stereotypes, misconceptions, and omissions about their history and identity.
Disabled People & Ableism	Allow for optimal self-identification and minimize the presence of ableism in the development of assessment materials. Some individuals with disabilities prefer to put the person before their 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 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.

Terminology	Description
Disabled People & Ableism, cont.	Avoid euphemisms such as “challenged.” Avoid using “click” for selecting answers and use more inclusive terms such as “select,” keeping in mind students who use assistive technology to interact with assessments. This also allows the item to be transferred more seamlessly between computer-based and paper-pencil environments. Remember that the term “neurodiversity” refers to the diversity of all people but 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 avoid 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,” “elderly adults,” “seniors,” or “senior citizens.”

A FINAL WORD

Examples of fair and unfair test materials are provided in Appendix A, and a Fairness and Sensitivity Quality Review Checklist is provided in Appendix B to help guide your qualitative evaluation of the fairness of assessment content.

Neither this nor any other set of guidelines can cover all the possible variations in content that will have to be evaluated for fairness in the Smarter Balanced assessments. When new issues arise that are not outlined in the *Fairness Guidelines*, consider the questions in the callout box below.

These three questions help guide decision making in any situation:

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?

From all of us at Smarter Balanced, we thank you for the time and care that you take in ensuring that our assessments are fair and equitable for all test takers. Please let us know if there is anything in these guidelines that can be improved to make the process easier for item developers and reviewers.

CITED 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.

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. Educational Assessment, 27(2), 170-178.

Ravitch, D. (2003). *The language police: How pressure groups restrict what students learn*. New York: Knopf.

Russell, M. (2024) Systemic Racism and Educational Measurement. Routledge.

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. (2024). *Smarter Balanced Usability, Accessibility, and Accommodations Guidelines*. Available from https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf

Smarter Balanced. (2012). *Smarter Balanced 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.

ADDITIONAL RESOURCES

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.

Crenshaw, K. W. (2017). *On intersectionality: Essential writings*. The New Press.

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.

APPENDICES

APPENDIX A: EXAMPLES OF FAIR AND UNFAIR TEST MATERIALS

The following examples are excerpts from test items and stimuli. Some of the excerpts illustrate various violations of the *Fairness Guidelines*. Others illustrate content that is fair in terms of the criteria provided in this document.

Want to try this out on your own? Print out this page and fold it in half. Read the prompts and see if you think they are fair or unfair, and why. Then turn over the page to see the comments and rationale provided.

Mathematics Content

The first set of examples consists of mathematics problems or excerpts from mathematics assessment content. These examples highlight the challenge of incorporating realistic context while minimizing linguistic complexity and avoiding irrelevant knowledge requirements.

Item Prompt	Evaluation
The drawing below shows seismic bracing. Highlight the two triangles that are congruent with each other.	**Unfair.** Few children are likely to be familiar with seismic bracing, and knowledge of seismic bracing is not related to the purpose of the question.
In the drawing below, highlight the two triangles that are congruent with each other.	**Fair.** The reading load is reduced, and there is no unfamiliar context.
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?	**Unfair.** The linguistic load is high for a mathematics 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. ("Perpendicular" is fair as a valid mathematical term.)
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?	**Fair.** 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.
A modem can send *x* bits per second. Write an expression that shows how many seconds it would take to send *y* bits.	**Unfair.** 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.

Item Prompt	Evaluation
Lee can walk *x* miles an hour. Write the expression that shows how many hours it would take Lee to walk *y* miles.	**Fair.** The context is familiar.
If one card is taken at random from a deck of playing cards, what is the probability that the card will be an ace?	**Unfair.** The question assumes knowledge of the number of aces and the total number of cards in a deck of playing cards. It is fair to ask about probability, and it is fair to use playing cards in mathematics problems. According to the guideline about gambling, however, it is not fair to assume that test takers have knowledge of the characteristics of a deck of playing cards.
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?	**Fair.** No knowledge of the characteristics of a deck of cards is required to answer the item.
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?	**Unfair.** 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.
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?	**Fair.** The reading load is reduced, and the context is familiar.
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	**Unfair** 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. **Fair** 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.

Item Prompt	Evaluation
According to the graph, the number of unemployed workers was highest in which year?	**Fair.** The mention of unemployed workers is fair.
Marisa hit the bull's eye with her rifle 7 times out of 9 shots. What percentage of the time did Marisa hit the bull's eye?	**Unfair.** 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 is controversial in any case.
A pizza is cut into 8 slices. If 5 students eat 1 slice each, how many slices will be left?	**Fair.** One slice of pizza is not excessive consumption of food.

English Language Arts (ELA) and Literacy

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

Item Prompt	Evaluation
Wagner used the orchestra to achieve certain effects in much the same way that other composers of operas used the singers.	**Fair** 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. **Unfair** if understanding the passage required knowledge of opera and how composers "used" the orchestra or "used" singers.
The African Americans living in Middletown tended to be part of households consisting of extended families living together.	**Fair.** The statement of fact about a particular group of African American people is fair and does not reinforce a stereotype.
Cyanide is one of the fastest-acting poisons known to science.	**Unfair.** The excerpt violates the guideline about avoiding dangerous actions and substances. Parents are likely to oppose including information about lethal substances in the test.
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.	**Unfair.** Excessive detail about the suffering of children makes the excerpt unfair.
Harlow was best known for the experiment in which he separated infant monkeys from their mothers shortly after the infants were born.	**Unfair.** The excerpt violates the guideline that prohibits inclusion of painful or harmful experimentation. The excerpt would be fair in a psychology test, however.
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.	**Unfair.** 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.
An ancestor of the modern horse the size of a dog gave rise to progressively larger species.	**Fair.** The passage concerns the evolution of horses. The guidelines identify the evolution of human beings as the aspect of evolution to avoid.
The Japanese immigrants enrolled in Ms. Kubota's class worked very hard.	**Fair.** The reference is to a particular group of Japanese immigrants, so it does not stereotype all Japanese immigrants.

Item Prompt	Evaluation
The amount of caffeine in a cup of coffee can still affect the human body more than three hours after it has been ingested.	**Fair.** The mention of caffeine appears to be in an objective discussion of the effects of drinking coffee and would follow the guideline on harmful substances if the passage did not encourage the drinking of coffee.
People who drive gas-guzzling SUVs contribute to global warming.	**Unfair.** The term "gas guzzling" is a colloquialism that students may not be familiar with. This topic is also controversial across states. The excerpt is unfair because it advocates for one side on a controversial topic.
In the 17th century, many convicted criminals were hanged, but some were slowly crushed to death.	**Unfair.** Death by slow crushing is clearly out of compliance with the guidelines about death and suffering.
A large influx of immigrants will destroy the equilibrium of a neighborhood.	**Unfair.** 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.
There has been an increase in the number of people who identify themselves as Native Americans.	**Fair.** "Native American" is appropriate and preferred over "American Indian." The fact that more people than before identify themselves as Native Americans is not a fairness problem.
Surprisingly, a girl won the math contest.	**Unfair.** By expressing surprise that a girl won the math contest, the excerpt reinforces the stereotype that girls have less quantitative ability than boys.
The soldiers and their wives attended the ceremony.	**Unfair.** 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.
… that all men are created equal, that they are endowed by their Creator with certain unalienable Rights . . .	**Fair.** Despite the use of "men" to refer to all people and despite the reference to God, the excerpt is fair because it is from the United States Declaration of Independence, an important historical document required by state standards.

Item Prompt	Evaluation
Bridges with steel frames are more likely to survive an earthquake than stone bridges.	**Fair.** The mention of a natural disaster is fair if there is no focus on death and destruction.
Lee's father and Juan's father are both policemen.	**Unfair.** Even though both officers are male, "police officers" is preferred to "policemen" to avoid the impression that only men are police officers.
The ancient Romans played handball and engaged in other sports while nude in the public baths.	**Unfair.** Though unintended, "engaged in other sports while nude" could be interpreted as a sexual innuendo.
The men's room is on the right; the girls' room is on the left.	**Unfair.** Parallel language would call for "women" to match "men" or "boy" to match "girl."
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 one tribe.	**Unfair.** 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 groups of people what they prefer to be called.
Frederick Douglass, the great African American abolitionist, was said to be born on Valentine's Day.	**Fair.** The excerpt requires no knowledge of how or why Valentine's Day is celebrated, nor any agreement that it should be celebrated.
Edward Said and Daniel Barenboim co-founded a children's orchestra.	**Unfair.** 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.
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."	**Fair.** Mention of slavery is fair, and state standards call for the inclusion of documents important in American history, such as the Emancipation Proclamation.

English Language Arts Prompts
The last set of examples are from ELA items and excerpts from items.

Item Prompt	Evaluation
According to the passage, how long ago did Homo sapiens evolve into a distinct species?	**Unfair.** 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.
The character delivering the monologue attributes the arrogance of the French to which of the following?	**Unfair.** Describing all 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.
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.	**Unfair.** The influence of people on climate change is controversial and out of compliance with the guideline on the avoidance of advocacy.
The author compares the artist's use of color to which of the following?	**Fair** if direct experience of color is not required to understand the passage and answer the items. **Unfair** if direct experience of color is required. The material would be unfair for students who are blind.
Our society stereotypes old people as weak, uninformed, forgetful, and foolish. Discuss the extent to which you agree or disagree with this stereotype.	**Unfair.** The question blatantly reinforces stereotypes of a group and invites test takers to agree with the offensive stereotypes.
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.	**Unfair.** The excerpt violates guidelines about the avoidance of religious material, even though the students are not asked to write directly about religion.
In the play, Luz was restricted to a wheelchair for which of the following reasons?	**Unfair.** The phrase "was restricted to a wheelchair" should be replaced with more objective terminology such as "began using a wheelchair."
According to the newspaper article, Robert died how many years after his brother John?	**Fair.** According to the *Fairness Guidelines*, it is fair to mention death if gruesome details are not depicted.

Item Prompt	Evaluation
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?	**Unfair.** 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.
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?	**Unfair.** Diseases that affect specific groups of people are likely to be problematic in terms of fairness.
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?	**Unfair.** 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.
The video excerpt of Baryshnikov dancing in *The Nutcracker* best illustrates which of the following aspects of his work described in the magazine article?	**Fair** if all the information needed to respond to the item is included in the video excerpt and the magazine article. **Unfair** if knowledge of ballet is required to answer the item. Only social dancing of couples is prohibited by the *Fairness Guidelines*.
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.	**Unfair.** Mention of slavery as a topic is fair, 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 students.

APPENDIX B:
FAIRNESS AND SENSITIVITY QUALITY REVIEW CHECKLIST

This checklist is meant to support test content review based on the ideas presented in the *Fairness and Sensitivity Guidelines*. It is not comprehensive but should be used as part of a larger review and discussion process during assessment review.

Topic:

- ☐ focuses on relevant content-standard skills and knowledge that are grade- and age-appropriate.
- ☐ is clearly presented with formatting and layout that emphasizes important content, including accessible graphics.
- ☐ is approached thoughtfully and with care if the topic is sensitive.
- ☐ includes all necessary contextual information in the prompt or stimulus.
- ☐ includes information that students with disabilities can access at an equitable level as their peers.
- ☐ contributes to the overall diversity of representation in the test content pool.
- ☐ avoids stressful, upsetting, or political topics

Language:

- ☐ is familiar to students and terminology that is frequently used in the classroom.
- ☐ is consistent throughout the stimulus and item (e.g., within the same item, avoid using “plant” in one instance and “flower” in another).
- ☐ avoids figures of speech, such as idioms and metaphors, unless required by the construct being tested.
- ☐ avoids jargon and colloquialisms, unless required by the construct being tested.
- ☐ avoids stereotyped language and culturally insensitive terminology.
- ☐ avoids complex grammar, such as compound sentences and passive voice.

This page is intentionally left blank.