# ETS Guidelines for Developing Fair Tests and Communications (2022)

# Table of Contents

# Foreword

The year 2020 was pivotal in many ways, especially when it came to the cultural transition of US society. Since then, many citizens have begun to collectively reckon with and reconsider their views about equity, fairness, and social justice. The resulting increased awareness of the systemic racism and of the profound inequities that have informed our history has the potential to be transformative.

With this awareness has come a more concerted effort by many Americans to reconsider how we should talk about social justice and, more specifically, about equity, diversity, and inclusivity. Along with the efforts to rethink and reconsider these issues has been the parallel goal to work toward changing fundamental social policies and practices in order to achieve greater equity for all groups and all individuals who live in this society.

Fairness has always been a central tenet of ETS products and services, as has been our commitment to continually challenge and evolve our understanding of its meaning. We are dedicated to participating in meaningful efforts to work toward social justice. To this end, the *ETS Guidelines for Developing Fair Tests and Communications* is an essential tool in accomplishing our organizational mission "to advance quality and equity in education by providing fair and valid assessments, research, and related services."

Reviews for the fairness of ETS materials have been carried out on a voluntary basis since the 1960s. The reviews became mandatory in 1980, when the first version of these written guidelines was issued. Since that time, we have updated the *Guidelines* approximately every five years. Our four decades of experience with the use of the *Guidelines* to ensure the fairness of our assessments and communications have helped to shape this most recent version.

As societal views of fairness have evolved, and as more has been learned about fairness, we have made the *Guidelines* increasingly inclusive and comprehensive. Notable updates to this edition, for example, include a broader treatment of gender and sexual orientation as well as the addition of a section on fairness in artificial intelligence algorithms. The 2022 edition of the *Guidelines* continues to recommend proactive representation of diverse racial, ethnic, gender, sexual orientation, and ability groups; to ensure that the pool of item writers and reviewers is as diverse as possible; and to provide guidance on current appropriate terminology for these groups. Note, however, that given the practical challenges posed by emerging definitions of fairness, this document will necessarily be a transitional one. That is, we fully realize that these recent revisions to this edition may well not be enough.

Traditional views of fairness were premised on the idea of equal treatment achieved in part through doing no harm to members of any given group by, for example, preventing bias from appearing in test materials with the use of such mechanisms as item-writing guidelines, differential item functioning analyses, and fairness reviews. Emerging voices within the educational measurement community, however, are increasingly recommending that assessments take a more proactive, specifically an antiracist, approach that directly addresses larger societal efforts to facilitate equity, including fair measurement in education, for all members of all groups.

For testing organizations like ETS, these efforts pose opportunities in the form of challenges. It is not fully clear *how* to practically implement assessments that reflect the recently emerging views about social justice nor how the fairness of such implementations might be evaluated. Yet, ETS is steadfast in its commitment to exploring and recommending solutions to such challenges and to continuing to innovate and adapt in service of our mission.

I am pleased to issue the 2022 edition of the *ETS Guidelines for Developing Fair Tests and Communications*. It is my intention that the *Guidelines* be updated on an ongoing basis as scientific research in assessment and societal changes influence views of fairness, equity, and social justice. In the interim, I hope that the *Guidelines* and the views of fairness expressed in the document will be of service not only to people at ETS but to all who are concerned about the fairness of tests and other communications.

Ida Lawrence

Senior Vice President, Research and Development

Educational Testing Service

# 1.0 Introduction

## 1.1 Purpose and Overview

The primary purpose of the *ETS Guidelines for Developing Fair Tests and Communications (GDFTC)* is to enhance the fairness, effectiveness, and validity of tests and test scores,[1] communications, and other materials created by Educational Testing Service (ETS). The *GDFTC* is also intended to help users do the following:

- better understand fairness in the context of assessment
- include appropriate content as materials are designed and developed
- avoid the inclusion of unfair content as materials are designed and developed
- find and eliminate any unfair content as materials are reviewed
- represent diversity appropriately in materials with an aim to increase inclusivity across all assessments as appropriate
- address issues related to accessibility and inclusion
- reduce subjective differences in decisions about fairness

To meet those purposes, we[2] do the following:

- We first describe the intended uses of the *GDFTC* and provide a rationale for its use in the design, development, and review of ETS materials.
- We then evaluate several definitions of the fairness of tests. The definition that forms the basis for the guidelines is that a test is fair if it is equally valid for the different groups of test takers affected by the test. We list the groups of people who should receive particular attention regarding fairness concerns.
- Next, we describe the various factors that affect the stringency or leniency with which you should apply the guidelines.
- We then list the basic principles for fairness in assessment to provide a basis for the detailed guidelines that follow.
- Then we discuss guidelines that focus on the avoidance of unnecessary barriers to the success of diverse groups of test takers. We include three types of barriers:

[1] We are aware that validity refers to the inferences and actions based on test scores rather than to the test itself, but for brevity in the *GDFTC* we will refer to the validity of a test and test scores or the validity of measurement.
[2] The compilers of the *GDFTC* will be referred to as “we,” and the readers will be addressed as “you.”

i. the measurement of knowledge, skills, or abilities unrelated to the purpose of the test

ii. the inclusion of material unrelated to the purpose of the test that raises strong negative emotions in test takers

iii. the presence of physical obstacles unrelated to the purpose of the test

- In addition to avoiding unnecessary barriers, fairness requires treating all test takers with respect. Important aspects of doing so that are discussed in the *GDFTC* include using appropriate terminology for groups and representing diverse people in test materials.
- The next section of the *GDFTC* includes additional guidelines for the fairness of the National Assessment of Educational Progress (NAEP) and for the fairness of K–12 tests.
- This is followed by guidelines for the fairness of artificial intelligence (AI) algorithms, which is followed by a very brief concluding section.
- Then we present a list of references, followed by a glossary of technical terms used in the document.
- Appendix 1 consists of information to help you use plain, easily understood language.
- Appendix 2 is an abridged list of the guidelines to use as a quick reference work aid once you have become familiar with the more detailed contents of the *GDFTC*.

### 1.2 Intended Uses

Although the focus of the *GDFTC* is on tests, the *GDFTC* applies to ETS products that include language or images in any medium. The principles for fairness described in it apply not only to tests but also to all ETS learning products and services and to all communications. All ETS material that will be distributed to 50 or more people outside of ETS must be reviewed for compliance with the *GDFTC*. The *GDFTC* includes a separate set of guidelines for developing and using ETS artificial intelligence (AI) systems.

Examples of ETS materials to which the *GDFTC* applies include, but are not limited to, artificial intelligence algorithms, books, cognitive and noncognitive tests, curricular materials, equating sets, formative tests, instructional games, interactive teaching programs, items (test questions) and stimuli, journal articles, learning products, news releases, photographs, pilot tests, posters, presentations, pretests, proposals, questionnaires, research reports, reviews, speeches, surveys, teaching materials, test descriptions, test-preparation materials, tests used in research studies, tutorials, videos, and Web pages.

Use of the *GDFTC* is not limited to ETS staff and associates. The *GDFTC* is copyrighted, but it is not confidential. The *GDFTC* will be useful to people—such as clients, potential clients, score users, and test takers—who are interested in how ETS strives to enhance the fairness of the materials it produces. Furthermore, ETS encourages the use of the concepts discussed in the *GDFTC* by all who wish to enhance the fairness of their own tests. To help make the *GDFTC* useful for people who are not familiar with the specialized vocabulary of testing, we have tried to avoid technical terms and have provided a glossary for the terms we need to use.

### 1.3 Reasons for Using the GDFTC

The main reason to use the *GDFTC* is that compliance with the guidelines will result in better ETS materials by helping you to do the following:

- ***Fulfill the ETS Mission.*** The ETS mission is, in part, "to advance quality and equity in education by providing fair and valid assessments, research, and related services." Because the *GDFTC* focuses on ways to enhance validity and fairness, its use supports the ETS mission.
- ***Meet Professional Testing Standards.*** According to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 63), "All steps in the testing process . . . should be designed in such a manner as to minimize construct-irrelevant variance [score differences not related to the purpose for giving the test] and to promote valid score interpretations." The *GDFTC* helps you to comply with the AERA, APA, NCME *Standards* because increasing fairness necessarily increases validity and reduces construct-irrelevant sources of score differences. The ETS Standards for Quality and Fairness[3] requires ETS to "follow guidelines designed to eliminate symbols, language, and content that are generally regarded as sexist, racist, or offensive, except when necessary to meet the purpose of the product or service" (ETS, 2014, p. 21). Such guidelines are provided by the *GDFTC*.
- ***Comply with Widely Used Editorial Policies.*** The *GDFTC* is consistent with the relevant sections of such commonly referenced sources for writers as the *Associated Press Stylebook* (Associated Press [AP], 2019), the *Chicago Manual of Style*, 17th edition (University of Chicago Press, 2017), and the *Publication Manual of the American Psychological Association*, 7th edition (APA, 2020).

[3] The *ETS Standards for Quality and Fairness (ETS Standards)* were initially adopted as corporate policy by the ETS Board of Trustees in 1981. They are periodically revised to ensure alignment with current measurement industry standards as reflected by the *Standards for Educational and Psychological Testing*.

### 1.4 When to Use the *GDFTC*

Several earlier editions of this document had the words “Fairness Review” in the title. We removed the word “Review” in more recent versions to avoid giving the impression that the guidelines were used only to check already developed materials. In fact, concern with fairness begins as materials are being designed. If there are several equally appropriate ways to measure a given topic, you should consider these guidelines and available evidence about group differences in scores in determining how best to measure it. For example, if a topic could be measured equally well with or without the use of complex graphs, decisions about the best way to measure the topic should take into account the fact that complex graphs may impede accessibility for people with certain disabilities. In general, if there are equally valid, equally practical, and equally appropriate ways to measure the same thing, preference should be given to the measures that result in smaller group differences in scores.

There are essentially two ways that lead to designing materials that are not fair:

- including the wrong content and skills
- failing to include a good sample of the right content and skills

Therefore, in addition to avoiding potentially unfair material during test design, it is very important to ensure that a good sample of the important content and skills is included. If groups of people differ, on average, in attainment of an important and relevant skill, then a test that fails to measure that skill would be less fair to the groups with higher attainment of that skill. For example, consider a subject in which writing skill is important. A combined direct measure of both actual writing and answering multiple-choice items would be fairer to a group that excels in writing than a multiple-choice test alone would be.

All people who develop materials for ETS or oversee scoring of ETS assessments should be trained to comply with the *GDFTC* to help avoid the inclusion of unfair content and to help ensure the inclusion of appropriate content. Waiting for the review stage to consider fairness is counterproductive and exposes ETS to the added time and expense of rework that could easily have been avoided by earlier attention to fairness. The reason for doing a review for fairness near the end of the process is to help ensure that the work done regarding fairness at the design and development stages was effective.

# 2.0 Meanings of Fairness for Tests

To make the types of judgments required to apply the guidelines properly, it is necessary to understand what is meant by fairness in the context of tests and related materials. Defining fairness for the purpose of these guidelines is challenging, however, because people have very different ideas about the meaning of fairness.

## 2.1 Definition Based on Common Usage

One of the difficulties in defining fairness in the context of assessment is that the common concept of fairness, including the perception of any inequity, is very broad. Fairness defined as any inequity can thus affect an individual as well as a group of people. For example, a younger sibling may say it is “unfair” that an older sibling is allowed a later bedtime. In a more germane context, students could say it is “unfair” for a teacher to include a question on a test about a topic that was never mentioned in class, even if every student in the class is affected in the same way. Many of the standards discussed in the document *ETS Standards for Quality and Fairness* address this broader concept of unfairness as being any inequity. While the *GDFTC* provides recommendations about how to promote diversity, representation, and equity in ETS products, the focus of the *GDFTC* is on unfairness caused by inappropriate content or images that adversely affect diverse groups of people, such as those described in the section of the *GDFTC* titled “Groups to Consider.”

## 2.2 Definition Based on Differences in Difficulty

Many people believe that items or tests that are harder for one group than for another group are not fair for the lower-scoring group. Although this belief that group score differences are in themselves proof of bias in tests is still widespread among the general public, this perception is misleading. The fact that there are group differences on a given assessment doesn’t mean that the test is itself biased (AERA, et al., 2014). At the same time, however, tests (and, more important, how scores on the test are used) may well reflect overall bias in educational opportunities—and in society itself.

A simple physical measurement example may be helpful in defining bias. Tape measures show that the average height of adults exceeds the average height of children. This is not evidence of bias in tape measures, because there is an actual difference between the heights of the two groups. Similarly, students who majored in mathematics in college generally get higher scores, on average, on the Quantitative Reasoning section of the GRE than do students who majored in English. The cause of the difference in scores is real differences in quantitative knowledge, skills, and abilities between math majors and English majors, not bias in the test.

The point is that group score differences cannot serve as proof of bias, because the test may be accurately reflecting real differences in what the test is intended to measure. Group score differences should be investigated to help ensure that they are not caused by bias, but the score differences by themselves are not proof that the test is unfair. However, if there is an equally valid way to measure a construct that results in smaller differences across groups, it is preferable to use that approach.

### 2.3 Definitions Based on Outcomes

Psychometricians have proposed several quantitative definitions of fairness based on the outcomes of using the tests.[4] Unfortunately, the definitions do not agree on the fairness of a test. Furthermore, the definitions based on outcomes are of little direct use in the design and development of tests, because the definitions cannot be applied until the completed tests are used.

### 2.4 Definition Based on Validity

Validity is the most important indicator of test quality. Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy and appropriateness of inferences and actions* based on test scores." Messick's invocation of "actions based on test scores" makes clear that how tests are used and the consequences they have affect the evaluative judgment of validity. That is, test use that causes negative consequences that have a disproportionate impact on a group must be justified by empirical evidence and a strong theoretical or logical rationale.

More recently, Kane (2013) defined validity as the extent to which the claims made about test takers on the basis of their scores are plausible and backed by logical and empirical evidence.

Whatever a test is intended to measure is known in the language of testing as a "construct." The construct consists of a mix of some body of knowledge, some set of skills, some group of abilities, or a cluster of some other attributes. This mix is often referred to collectively as "KSAs."

Validity can be thought of as the extent to which a test measures a suitable sample of the important construct-relevant KSAs and minimizes the measurement of any construct-irrelevant KSAs. Perfect validity is impossible, but as the proportion of the score differences caused by important, construct-relevant KSAs increases, validity increases.[5]

Validity is directly tied to the degree to which test material is well chosen and construct relevant. Fairness is directly tied to the degree to which test material is equally valid for different groups of test takers. If a poorly chosen sample of content or construct-irrelevant KSAs affects all test takers to about the same extent, validity is diminished. If a poorly chosen sample of content or construct-irrelevant KSAs affects some group of test takers more than some other group of test takers, then both fairness and validity are diminished.

---

[4] Interested readers should refer to Cleary, 1968; Cole, 1973; Linn, 1973; Thorndike, 1971.

[5] Perfect validity is impossible, because some construct-irrelevant sources of score differences are always present, such as luck in guessing the correct answer to a question.

For test designers, developers, and reviewers the most useful definition of fairness in assessment is based on validity. Fairness is essential for validity and validity is essential for fairness. Shepard (1987, p. 179) very concisely defined bias as “invalidity.” According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014, p. 49), “fairness is a fundamental validity issue.”

Therefore, fairness in the context of assessment can usefully be defined as the extent to which inferences and actions based on test scores are equally valid for a diverse population of test takers.

The extent to which other products, services, and publications meet their intended purposes for their intended users is analogous to validity in tests. For example, educational products and services should increase the knowledge, skill, or other relevant abilities of the people who use them. The extent to which products, services, and publications meet their intended purposes for different groups of intended users is analogous to fairness in tests.

Material identified as inappropriate for tests is also likely to interfere with the effectiveness and fairness of other ETS products, services, or publications. For example, language that is more difficult than necessary to meet the purpose of a test or of a lesson will make the test less fair and the lesson less effective.

## 3.0 Groups to Consider

Ideally, the *GDFTC* applies to all groups of people, but special attention should be paid to groups that are discriminated against based on characteristics such as the following:

- age
- appearance
- citizenship status
- disability
- ethnicity
- gender (including gender identity or gender representation)
- national or regional origin
- native language
- race
- religion (or absence of religion)
- sexual orientation

- socioeconomic status

When evaluating fairness, it is also necessary to consider intersectionality. Intersectionality is a framework for understanding the ways in which intersecting identities (e.g., race and gender) factor into the experience of those who hold multiple marginalized identities (Crenshaw, 1991). For example, Black women may experience test material differently than do either Black men or White women.

# 4.0 Interpreting the Guidelines

This document contains flexible guidelines, not strict rules. For many of the guidelines, compliance is a matter of degree rather than a clear binary decision. The primary goal of using the guidelines is to increase the validity, effectiveness, and fairness of ETS products and services. What level of language and what content would best achieve those goals? At what point does the difficulty of language become a construct-irrelevant barrier to success? How controversial does content have to be to violate a guideline?

The *GDFTC* cannot eliminate all subjectivity. Material that seems acceptable to some may be rejected by others. How important for validity does content have to be to justify its inclusion if it appears to be out of compliance with a guideline? Subject-matter experts may disagree about the importance of certain content.

Judgment is required to interpret the guidelines appropriately. No compilation of guidelines can anticipate all possible circumstances and remain universally applicable without interpretation. It is important, however, to guard against both interpretations that are too weak and interpretations that are too stringent.

An overly lax interpretation of the guidelines may allow unfair content into ETS tests and reduce validity. On the other hand, an overly fervent application of the guidelines may interfere with validity and authenticity. Excessively zealous interpretations may also lower confidence in the value of the guidelines. The individual guidelines must be applied conscientiously, but with common sense, with regard to the client’s or user’s requirements, and with an awareness of the need to measure important aspects of the intended construct with realistic material.

The interpretation of the guidelines should vary with a number of factors. Consider the following factors when deciding whether material complies with the guidelines.

## 4.1 Importance for Validity

In deciding whether or not material complies with the guidelines, consider whether or not the material is important for valid measurement.

Because of the close link between validity and fairness, any material that is important for valid measurement—and for which a similarly important but more appropriate substitute is not available—may be acceptable for inclusion in a test, even if it would otherwise be out of compliance with the guidelines.

### 4.2 Need for Specific Content

Interpret the guidelines more stringently for items that primarily measure skills than for items that primarily measure content. Skills (e.g., collaboration, critical thinking, mathematical reasoning, reading comprehension, speaking, writing) can be applied across many different content areas. The valid measurement of skills seldom requires material that is out of compliance with any of the guidelines.

Items that primarily assess subject matter (e.g., biology, English literature, history, nursing, psychology) require that specific content be included for valid measurement. Some offensive or upsetting material may be important for validity in certain content areas. For example, detailed descriptions of the symptoms of certain illnesses would be potentially upsetting to test takers and would not be appropriate in a test of K-12 reading comprehension. However, this same content may be important for validity in a test for doctors, so it would be fair in that test. Along the same line, a United States history test may appropriately include detailed descriptions about the fatalities in the Vietnam War that would otherwise be out of compliance with the *GDFTC*. If it is important to measure the ability to compare and contrast different points of view about a topic, the topic must be controversial enough to allow at least two defensible points of view.[6] Similarly, noncognitive items may require certain content to measure attitudes, feelings, beliefs, interests, personality traits, and the like. For example, the only way to measure attitudes about abortion is to ask questions related to abortion.

### 4.3 Consequences of Using the Material

Some tests are used to help make high-stakes decisions about test takers (e.g., award of a high school diploma, college admissions, occupational licensing). Because such tests have important consequences for test takers, the guidelines should be interpreted strictly.

When the results of testing are less consequential, the guidelines may be interpreted more freely. For example, some tests do not report scores for individuals. Material used for instruction outside of a testing situation still needs to comply with the guidelines, but the guidelines may be interpreted more freely than in a testing context. The guidelines may be interpreted most freely for material that is discussed in class with the guidance and support of a teacher and the opportunity for students to ask questions. Instructional material designed for use without the support of a teacher needs to conform more closely to the guidelines, but not as closely as material used in a test that has important consequences for the test taker.

[6] If controversial content is included in a test because that material is important for valid measurement, ETS or the client may wish to indicate that it does not endorse the views expressed.

### 4.4 Age and Experience of Test Takers

Interpret the guidelines most strictly for younger test takers. (Additional guidelines for younger test takers are in the section on NAEP and K–12 testing.) In general, the older and more widely experienced the test takers are, the more freely the guidelines should be interpreted.

Consider the kinds of material that test takers are likely to have been exposed to when deciding whether some test material is likely to offend or upset them. If test takers have become accustomed to the material through repeated exposure in their studies, their occupations, or their daily lives, it is not likely that encountering it again in a test would be excessively problematic.

### 4.5 Control Over the Material

ETS has much more control over the material that it writes than it has over previously published material. Therefore, interpret the guidelines more strictly for original ETS material than for material from other sources. Clients may require the use of unedited, authentic materials as stimuli (e.g., excerpts from published documents, graphs, maps, photographs) in tests. Publishers or authors may forbid the revision of copyrighted materials.

A construct-relevant use of historical, literary, or other authentic materials published before current conventions about appropriate language were in place may result in apparent conflict with the guidelines. The use of such materials is acceptable if the use of unrevised, authentic materials is an important aspect of the intended construct or is required by the client, and if an effort has been made to obtain materials that minimize departures from the guidelines.

### 4.6 Directness of the Material

It is possible to create a far-fetched scenario in which any innocuous topic could be considered upsetting. For example, a reviewer could say that a picture of a mother and child would be upsetting to orphans. Items and stimuli about innocuous topics are generally acceptable, even if an atypical scenario could be constructed in which they might be upsetting. Contexts that directly mention an upsetting experience are less likely to be acceptable. For example, a mathematics item about the average speed of a car should not be construed as potentially upsetting for test takers who have been involved in a car accident. On the other hand, an item about the average number of children killed per year in car accidents would be unacceptable, unless it were important for validity and no similarly important substitute were available.

### 4.7 Extent of the Material

A brief mention of a problematic topic may be acceptable even though a more extended, detailed discussion of the topic should be avoided. For example, a statement that a cat killed a wild bird might be acceptable, but an extended, graphic description of the process would probably not be acceptable unless it were construct relevant.

### 4.8 Client Preferences

Different clients may reasonably have different opinions about what is considered fair. For example, one client may believe that references to social dancing in a K–12 test are acceptable, and another client may prefer to avoid the topic. One client may decide that the use of "Latinx" is appropriate, and another client may prefer to avoid that term.

Follow the fairness requirements of the client on such matters of opinion, but avoid departures from the guidelines that would clearly result in negative consequences for test takers such as the use of material that condones or incites hatred or contempt for people based on such attributes as culture, disability, gender, race, religion, or sexual orientation.

### 4.9 Country for Which the Test Is Designed

The *GDFTC* applies as written to materials designed primarily for use in the United States, even if the tests are administered worldwide. Materials designed specifically for use in other countries will very likely require changes in the interpretation of some of the guidelines and revisions to other guidelines. For example, even though the need to avoid material that is unnecessarily offensive to test takers is universal, exactly what is considered offensive will vary from country to country.

## 5.0 Principles and Guidelines for Fairness

Though it is possible for reasonable people to disagree about the value of certain guidelines, there are general principles for fairness that appear to be indisputable if fair measurement is a goal. In particular, tests and test items should do the following:

- Measure the important aspects of the intended construct.
- Avoid construct-irrelevant barriers to the success of test takers.
- Provide assessment design, content, and conditions that help diverse test takers show what they know and can do so that valid inferences are supported.
- Provide scores that support valid inferences about diverse groups of test takers.

The following sections contain specific guidelines, grouped by major topics, that support these general principles. If there is a disagreement about the interpretation of a guideline, follow the interpretation that best supports the general principles for fairness in assessment.

## 6.0 Construct-Irrelevant KSA Barriers to Success

Construct-irrelevant KSA barriers to success may arise when construct-irrelevant KSAs are required to answer an item correctly. Because of differences in environments, experiences, interests, and the like, different groups of people may differ in average knowledge of various topics and in average levels of various skills or abilities. If a construct-irrelevant KSA is required to answer an item, the validity of the item is diminished. If the KSA is not equally distributed across groups, then the fairness of the item is diminished as well.

For example, if an item that is supposed to measure multiplication skills asks for the number of meters in 1.8 kilometers, knowledge of the metric system is construct irrelevant. If, on average, one group of test takers is less familiar with the metric system than are other groups of test takers, the item would be less valid for one of the groups and, therefore, less fair. (Note: If no group on average had less knowledge of the construct-irrelevant content, the item would be fair, but it would be invalid.)

If, however, the intended construct were conversion within the metric system, then the need to convert kilometers to meters would be relevant to the construct and, therefore, valid and fair. Whether a KSA is important for valid measurement or is a source of construct-irrelevant differences depends on the intended construct.

Among common construct-irrelevant sources of KSAs are unfamiliar contexts, the effects of certain disabilities, unnecessarily difficult language, regionalisms, religion, specialized knowledge of various topics, translation, unfamiliar item types, and topics specific to the United States.

### 6.1 Contexts

In items that are intended to measure skills rather than specific content, stimuli, such as reading-comprehension passages, still have to be about something. Similarly, applications of mathematics usually require some real-world setting. The content of reading passages and the settings of mathematics problems have raised fairness issues. It is not appropriate to assume that all test takers have had the same experiences. What construct-irrelevant contexts are fair to include in tests?

In short, the answer depends on what test takers are expected to know about the context and on the extent to which the information necessary to understand the context is available in the stimulus material. Generally, school-based experiences are more commonly shared among students than are their home or community-based experiences.

When selecting contexts, strive to find contexts that are not only familiar but also appealing to different groups of test takers. Contexts should engage test takers rather than puzzle or distract them.

A very important purpose for reading is to learn new things. It could severely diminish validity to limit the content of reading passages to content already known by test takers. If the construct is reading comprehension rather than knowledge of the subject matter from which the passage is excerpted, then the construct-irrelevant information required to answer the items correctly should either be common knowledge among the intended test takers or be available in the passage. Similarly, for mathematics problems, the contexts should be common knowledge among the intended test takers, or the necessary information should be available in the problem. The teachers of students at the relevant grades are a very helpful source of information about what is considered common knowledge at those grades.

### 6.2 Disabilities

Do not use test items in which a correct response requires personal experiences that may be unavailable to test takers with disabilities, unless the item is required for valid measurement. For example, a test taker who uses a wheelchair can still understand a reading passage about a footrace, but a test taker who is deaf might have difficulty with a reading item related to phonics. A pie chart that provides its data only through the use of color would be problematic for test takers who are visually impaired. (Disabilities that affect a test taker's ability to see or hear test materials are discussed in the section titled "Construct-Irrelevant Physical Barriers.")

### 6.3 Language

Use the simplest and clearest language that is consistent with valid measurement. While the use of simple and clear language is particularly important for test takers who have limited English skills or language-related disabilities, the use of plain language is beneficial for all test takers when linguistic competence is construct irrelevant. Appendix 1 ("Plain Language") provides information about the use of easily understood language. Note that while "simple and clear" remains an important goal for communications, a more flowery approach may occasionally be appropriate for certain purposes.

Avoid requiring knowledge of the jargon or specialized vocabulary of an occupation or academic discipline unless such vocabulary is important to the construct being assessed. What is considered excessively specialized requires judgment. Take into account the maturity and educational level of the test takers when deciding which words are too specialized. Even if it is not necessary to know a difficult, construct-irrelevant word to answer an item correctly, the word may intimidate test takers or otherwise divert them from responding to the item.

Avoid requiring construct-irrelevant ability to interpret figurative language (e.g., hyperbole, idiom, metaphor, metonymy, personification, synecdoche) to answer an item correctly.

You should use difficult words and language structures if they are important for validity. For example, difficult words may be appropriate if the purpose of the test is to measure depth of general vocabulary or specialized terminology within a subject-matter area. It may be appropriate to use a difficult word if the ability to infer meaning from context is construct relevant. Figurative language may be appropriate in a language arts test. Complicated language structures may be appropriate if the purpose of the test is to measure the ability to read challenging material.

### 6.4 Regionalisms

Do not require knowledge of words, phrases, or concepts more likely to be familiar to people in some regions of the United States than by people in other regions, unless it is important for valid measurement. When there is a choice, use generic words rather than their regional equivalents. For example, more test takers—particularly those outside of the United States—are likely to understand the generic word “sandwich” than are likely to understand the regionalisms “grinder,” “hero,” “hoagie,” or “submarine.” Names used for political jurisdictions, such as “borough,” “province,” “county,” or “parish,” vary greatly across regions. Do not require knowledge of their meaning to answer an item unless such knowledge is part of the construct. Regionalisms may be particularly difficult for test takers who are not proficient in English and for young test takers.

### 6.5 Religion

Do not require construct-irrelevant knowledge about any religion to answer an item. If the knowledge is part of the construct, use only the information about religion that is important for valid measurement. For example, much European art and literature is based on Christian themes, and some knowledge of Christianity may be needed to answer certain items in those fields. Items about the religious elements in a work of art or literature, however, should focus on points likely to be encountered by test takers as part of their education in art or literature, not as part of their education in religion.

### 6.6 Specialized Knowledge

Avoid requiring construct-irrelevant specialized knowledge to answer an item correctly unless the test is structured to give examinees with different funds of specialized knowledge the opportunity to use that knowledge appropriately. For example, knowing the number of players on a soccer team would be construct relevant on a licensing test for physical education teachers, but it would not be construct relevant on a mathematics test.

What is considered specialized knowledge will depend on the education level and experiences of the intended test takers. Teachers of the appropriate grades, reading lists from various schools, vocabulary lists by grade, and content standards can all help determine the grades at which students are likely to be familiar with certain concepts.

The following subjects are likely sources of construct-irrelevant knowledge. Aspects of the subjects that are common knowledge and that the intended test takers are expected to be familiar with are acceptable. Do not, however, require specialized knowledge of these subjects unless that knowledge is construct relevant. The subject areas that require extra caution regarding specialized knowledge include, but are not limited to, the following:

- agriculture
- construction
- finance
- fine arts

- law
- medical topics
- military topics, weapons
- politics
- science
- sports
- technology
- tools
- transportation

For example, even if the test takers are adults, do not assume that almost all will have construct-irrelevant knowledge of words such as “combine” (as in “a combine harvester”), “joist,” “margin call,” “aria,” “subpoena,” “stenosis,” “filibuster,” “lumen,” “bunt,” “buffer,” “chuck,” and “sloop.” At the appropriate grade levels, however, almost all are likely to know the meaning of “tractor,” “board,” “bank,” “song,” “judge,” “skull,” “senator,” “thermometer,” “ball,” “computer,” “hammer,” and “boat.” The use of construct-irrelevant specialized knowledge will decrease validity in all cases, and it will specifically also decrease fairness if the specialized knowledge is not evenly distributed across diverse groups.

### 6.7 Translation

Translating test items without also accounting for cultural differences is a bad practice and a common source of barriers to success related to measurement of irrelevant knowledge. Translation alone may be insufficient for many test items. The content of items must be adapted for the culture of the country in which the items will be used. For example, an item in a test originally made for use in the United States could refer to the Fourth of July, a holiday that may not be familiar to test takers in other countries. Even in the absence of cultural differences, translation may change the difficulties of items because words that are common in the source language may be translated to words that are more specialized in the target language, and vice versa. Translation issues may exist even if the same language is used in various countries. For example, if tests are given in English, differences between American and British English in vocabulary and spelling may be a source of construct-irrelevant knowledge.

### 6.8 Unfamiliar Item Types

Technology-enhanced item types have many advantages, but they have also increased the amount of information and skill that test takers need in order to respond to the items. They have also increased the different types of items test takers interact with. For example, test takers may have to select cells in a matrix, construct graphs, or highlight sentences in a passage.

Test takers may have to use a mouse, trackball, or other device to “drag and drop” or use a keyboard, speech-to-text software, or another mechanism to enter a lengthy text response. Lack of the necessary information, poor design, or skill can be a construct-irrelevant barrier to success.

Assessments may also utilize audio and/or video stimuli in interactive, scenario-based tasks that require test takers to engage in simulations. The use of these types of technology should be reviewed to make sure that construct-irrelevant variance is not introduced and that the task type and functionality follow accessibility and best practices for inclusive design.

Make clear to the test taker what action is needed in order to respond to the item. Using the digital device should not be a construct-irrelevant source of difficulty. Make sure the item measures the intended construct, not the ability to use the digital device interface to respond to the item.

Be consistent in the way that the same or very similar items are presented. Avoid needless variation and construct-irrelevant complexity.

Clearly distinguish among items that appear to be similar but require different types of responses. For example, some multiple-choice items allow only a single response, and other multiple-choice items require selecting all of the responses that are correct.

### 6.9 United States Dominant Culture

ETS tests are taken in many countries. Even tests administered in the United States may be taken by newcomers to the country. Therefore, do not require a test taker to have specific knowledge of dominant United States cultures or conventions to answer an item, unless the item is supposed to measure such knowledge. For example, do not require knowledge of United States coins if the purpose of an item is to measure arithmetic, unless the construct includes knowledge of United States coins.

Unless it is part of the construct, do not require knowledge specific to the United States regarding topics such as the following:

- brands of products
- celebrities
- corporations
- culture
- customs
- educational systems
- elections

- food
- geography
- government agencies
- history
- holidays
- institutions
- laws
- measurement systems (inches, pounds, degrees Fahrenheit)
- money
- organizations
- pets (e.g., dogs are considered pets in some cultures, unclean in some other cultures, and food in still other cultures)
- places
- plants
- politics, politicians, political parties, political systems
- political subdivisions (local, state, federal)
- public figures
- regional differences
- slang
- sports and sports figures
- television shows and other entertainment
- wildlife

Do not assume that all test takers are from the United States. In general, it is best not to use the word “America” or the phrase “our country” to refer solely to the United States of America, unless the context makes the meaning clear. Similarly, unless the context makes it clear, do not use the phrase “our government” to refer particularly to the United States government. Popular names of places such as “the South,” “the Sun Belt,” “the Delta,” or “the City” should not be used without sufficient context to indicate what they refer to.

Some images or descriptions of people and their interactions that are acceptable in the United States may be offensive to people in certain other countries with conservative cultures. In tests that will be used worldwide, avoid construct-irrelevant images of people posed, dressed, or behaving in a way that may be perceived as immodest in another culture. Certain hand signals that are acceptable in the United States have offensive meanings in some other countries. For example, unless they are construct relevant, avoid images of gestures such as the OK sign (thumb and first finger forming a circle, other fingers extended) and the victory sign (first two fingers extended and spread apart, other fingers clenched).

Illustrations that are intended to aid understanding may be a source of construct-irrelevant difficulty if the depictions of the people do not meet the cultural expectations of test takers in countries other than the United States. People intended to be professors, for example, should look older than the students depicted and should be dressed conservatively.

## 7.0 Construct-Irrelevant Emotional Barriers to Success

Construct-irrelevant emotional barriers to success arise when language, scenarios, or images cause strong emotions that may interfere with the ability of some groups of test takers to respond to an item. For example, offensive content may make it difficult for some test takers to concentrate on the meaning of a reading passage or the answer to a test item, thus serving as a source of construct-irrelevant differences. Test takers may be distracted if they think that a test advocates positions counter to their strongly held beliefs. Test takers may respond emotionally rather than logically to excessively controversial material.

In determining whether or not test material could cause a construct-irrelevant emotional barrier, keep in mind that test takers may be anxious and may be feeling time pressure as they interact with test material. Therefore, avoid any construct-irrelevant material that may plausibly cause a negative reaction under those conditions, even if the content might appear to be balanced and acceptable based on a careful, objective reading in more comfortable circumstances. For example, the author of a reading passage may present both sides of a controversial issue, yet the inclusion of a position that some test takers strongly oppose may be an emotional barrier for them, regardless of the remainder of the passage. Also, avoid potentially offensive answer choices in multiple-choice items. Even an offensive wrong answer choice may be problematic, because a test taker who chooses it presumably believes that it is correct and that it represents the view of the author (and, arguably, the view of ETS).

Materials about a group of people who have been the object of discrimination need careful scrutiny for any construct-irrelevant content that might plausibly cause a negative reaction among members of the group. Avoid materials that depict painful current or past occurrences when there is no need to include the depiction for valid measurement. If the passage is about a group other than your own, you might find it helpful in evaluating the passage to consider how you would react if the passage were about a group to which you belong. Often you will need to make a special effort to understand that what may not at first seem problematic to you might in fact be problematic for others. No group of test takers should have to face material that raises strong negative emotions among members of the group, unless the material is important for valid measurement.

It is preferable, but not required, that passages about groups whose members have historically been discriminated against be written by a member of the group or represent the views of members of the group. In general, the authors of passages and the writers and reviewers of items should reflect the diversity of the population. Likewise, programs should proactively seek internal and external test developers who represent as diverse a population as possible and test developers should strive to find passages written by as diverse a population as possible.

### 7.1 Topics to Avoid

Some topics are so likely to cause negative reactions among test takers that they are best avoided in test materials unless they are important for validity. Some topics may be problematic simply from a public relations point of view.

Regardless of its inclusion in the following list, any topic that is important for validity, and for which there is no similarly important substitute, may be tested.

Any list of topics to avoid can be only illustrative rather than exhaustive. Current events, such as a highly publicized terrorist attack, a pandemic, or a destructive natural disaster, can cause new topics to become distressing at any time, so reviewers must always keep in mind recent controversies and other potentially upsetting events. Often, programs will want to search for key words in extant item pools to make sure that items that were previously deemed acceptable are not now problematic because of recent events. A topic is not necessarily acceptable merely because it has not been included in the following discussion. Therefore, it is a good practice to obtain a fairness review of any potentially problematic material before time is spent developing it.

Unless they are important for validity, avoid topics that are as likely as the following to trigger negative reactions:

- abduction
- abortion
- abuse of people or animals
- acquiring a disability or serious disease
- alcoholism
- amputation
- atrocities

- blasphemies, curse words, obscenities, profanities, swear words, vulgarities
- bodily gases, bodily fluids, bodily wastes
- bullying
- cannibalism
- civil protests or riots
- contraception
- discrimination
- disfigurement
- drug addiction, drug overdose, drug use
- eating disorders
- eugenics
- euthanasia
- forced migration
- forced quarantine
- genitalia
- genocide
- gruesome, horrible, or shocking aspects of accidents, deaths, diseases, natural disasters, or other causes of suffering
- home invasion
- homophobia and transphobia
- human trafficking
- hunting or trapping for sport
- incest
- murder
- mutilation
- painful or harmful experimentation on human beings or animals
- pandemics (epidemics, plagues, viruses, contagions, quarantine, vaccine)

- pedophilia
- racial or ethnic (including White) supremacy or difference
- rape or sexual assault
- Satanism
- starvation
- suicide, self-harm, self-destructive actions
- terrorism
- torture
- unsafe activities
- war
- witchcraft

### 7.2 Topics Requiring Care

While some topics may not necessarily trigger negative reactions, they need to be treated in as balanced, sensitive, and objective a manner as is consistent with valid measurement.

***Advocacy.*** Items and stimulus material should be neutral and balanced whenever possible. Do not use test content to advocate for any contested cause or ideology or to take sides on any controversial issue unless doing so is important for valid measurement. Test takers who have opposing views may be disadvantaged by the need to set aside their beliefs to respond to items in accordance with the point of view taken in the stimulus material.

Some types of items, such as the evaluation of an argument, require the presentation of a particular point of view, however. Such items should be no more controversial than is necessary for valid measurement. Communications other than test materials may advocate for those causes on which ETS has taken a position.

***Avatars.*** In some scenario-based items, the test takers use avatars to represent themselves or other characters on a digital device. If realistic avatars are used, the mix of genders, races, and ethnicities should comply with the section of the *GDFTC* titled "Representation of Diversity." Be careful to avoid reinforcing stereotypes in depicting avatars that represent various groups. One possible strategy to avoid diversity concerns is to use unrealistic, cartoonlike avatars that do not represent any identifiable gender, race, or ethnicity. Note that some characters (e.g., animals) will be differentially familiar and may have different associations, depending on culture and country. Note that design and functionality of avatars must comply with accessibility and best practices for inclusive design.

***Biographical Material.*** Avoid items or stimuli that focus on individuals who are associated with offensive topics or controversial activities unless the use of such items or stimuli is important for valid measurement. If an item mentions a real person who is unknown to you, consult colleagues or reference materials to determine whether the person is associated with inappropriate topics or activities. Unless important for validity, avoid biographical passages that focus on live celebrities, whose future actions are unpredictable and may result in fairness problems.

***Brand Names.*** Avoid construct-irrelevant brand names, because the mention of a brand in a positive or even a neutral context could be taken as advocacy for the product. Mention of the brand name in a negative context could be construed as a criticism of the brand. Be careful to avoid brand names even when the brand name has become better known than the generic name for a product (e.g., Band-Aid for adhesive bandage, Vaseline for petroleum jelly, Kleenex for facial tissue, or Google as a transitive verb for searching the Web). Communications other than test materials may mention brands as appropriate.

***Conflicts.*** Unless important for validity, do not take the point of view of one of the sides in a conflict in which test takers may sympathize with different factions. Do not focus on prominent participants in the conflict. One side's courageous freedom fighter is the other side's cowardly terrorist. In particular, the material should not appear to be propaganda for one of the sides in the conflict if there are test takers who may favor the other side.

***Cryptic References.*** Materials used in tests come from many sources. Some of those sources may contain cryptic references to anti-Semitism, drugs, gangs, homophobia, racism, sex, White supremacy, and other unsuitable topics. Be alert for such references and try to avoid them in tests unless they are important for validity.

Some cryptic references substitute numbers for letters (1 = A, 2 = B, etc.). For example, the number 88 is used to stand for "Heil Hitler." The number 311 (three times K, the 11th letter) is used to stand for "Ku Klux Klan." Other cryptic numbers come from various sources. For example, the number 666 is associated with Satanism, the number 14 and the phrase "14 words" are associated with a White supremacist slogan, the date April 20 is Hitler's birthday, and the time 4:20 and the number 420 have become associated with drug use.

Some apparent nonsense syllables that might be disguised as names of fictitious people or places have hidden meanings. For example, "akia" stands for "A Klansman I am," the word "orion" stands for "our race is our nation," and the word "rahowa" stands for "racial holy war."

Cryptic references (such as pictures of people flashing gang or White supremacist hand signs) to inappropriate topics can be embedded in images or symbols. Many seemingly innocuous images (e.g., eggplant, peach) may have sexual meanings in the world of sexting emojis. Refer to the section of the *GDFTC* titled "Visual Material."

Cryptic references can be a problem because there are so many and because they change so rapidly, so test developers are likely to be unaware of all of them. Use search engines such as https://www.adl.org/hate-symbols to check possible cryptic references to hate groups, such as names, numbers, images, or words that look odd, out of place, or unnatural or that appear to be arbitrary.

***Disability.*** Avoid negative or derogatory references to people with disabilities. Avoid the implication that people with disabilities are less valuable members of society than are members of the general population. People with disabilities should be represented in test materials as described in the section of the *GDFTC* titled "Representation of Diversity."

***Evolution.*** The topic of evolution has caused a great deal of controversy. The most sensitive aspect of evolution appears to be the evolution of human beings. Therefore, avoid items or stimuli concerning the evolution of human beings and the similarities of human beings to other primates unless such test content is important for valid measurement. Any aspect of evolution is allowed if it is important for valid measurement.

For K–12 tests, the jurisdictions that commission the tests control the contents of their tests. Some states restrict any mention of evolution in skills tests. Some states also restrict topics associated with evolution, such as dinosaurs, fossils, or the age of Earth. Please refer to the section of the *GDFTC* titled "Additional Guidelines for Fairness of NAEP and K–12 Tests" for more information.

***Group Differences.*** Avoid unsupported generalizations about the existence or causes of group differences. Do not state or imply that any groups are superior or inferior to other groups with respect to such traits as caring for others, courage, honesty, trustworthiness, physical attractiveness, or quality of culture. Do not overrepresent members of a group as showing irrational or criminal behavior.

Do not treat any one group as the standard of correctness against which all other groups are measured.[7] For example, the phrase "culturally deprived" implies that the dominant culture is superior and that any differences from it constitute deprivation.

***Humor, Irony, and Satire.*** Avoid construct-irrelevant humor, irony, and satire, because people may not understand them or may be offended or distracted by them. People with certain cognitive disabilities may have difficulty understanding them. In particular, avoid construct-irrelevant humor, irony, or satire that is based on disparaging any group of people, their culture, their strongly held beliefs, or their concerns. It is acceptable to test understanding of humor, irony, and satire when it is important for valid measurement as in, for example, the interpretation of a political cartoon in a social sciences test.

[7] This does not apply to norm groups used in score reporting or reference groups used in statistical analyses.

***Luxuries.*** Avoid depicting situations that are associated with excessive spending on what some members of the test-taking population would consider luxuries (e.g., cruises, designer clothing, private swimming pools, vacation homes), unless the depiction is important for validity. The goal is to avoid making many test takers feel excluded by unnecessarily depicting activities and material goods associated with the wealth of a small percentage of test takers.

***Maps.*** Unless important for valid measurement, avoid showing maps of politically disputed areas indicating that the area belongs to one of the parties in the dispute.

***Mistreatment of Groups.*** Unless it is important for validity, avoid material that focuses on any group that has been the object of discrimination if the group is depicted as

- passively suffering the effects of prejudice;
- being harmed, exploited, or subjected to cultural appropriation by a supposedly superior group;
- being improved by contact with a supposedly superior group; or
- emulating a supposedly superior culture.

The goal is to avoid upsetting members of the group depicted in the materials. Therefore, a brief mention of an issue of concern in materials that are clearly focused on an unobjectionable topic may be acceptable.

***Personal Questions.*** Avoid asking test takers to respond to excessively personal questions regarding themselves, their family members, authority figures, or their friends. Questions about topics such as the following are inappropriate unless important for validity or required for determining qualification for some program or benefit:

- antisocial, criminal, or demeaning behavior
- citizenship
- disability
- family or personal wealth
- general health
- political party membership
- mental health
- relationship status
- religious beliefs or practices or membership in religious organizations
- sexual orientation, practices, or fantasies

***Religion.*** Avoid construct-irrelevant material that focuses on any religion, any religious group, any religious holidays, any religious practices, any religious beliefs, any conflicts between religions, or anything closely associated with religion (including the creation stories of various cultures) unless it is important for valid measurement. Also avoid material on the lack of religion, agnosticism, or atheism.

Brief references to religion, religious roles, institutions, or affiliations are acceptable as long as they do not dwell on the subject of religious beliefs and practices. For example, a passage on Japan may indicate that Shinto and Buddhism are the country's two major religions. A passage on Dr. Martin Luther King, Jr., may indicate that he was a minister or that he worked with the Southern Christian Leadership Conference.

Do not support or oppose religion in general or any specific religion or lack of religion. Do not praise or ridicule the practices of any religion. Try to avoid using phrases closely associated with religion as figures of speech (e.g., "born-again" as a general intensifier, "cross to bear" to stand for a person's problem). It is generally preferable not to use the words "crusade" or "crusader" outside of their historical context, although there might be reasonable exceptions (e.g., a reference to James Bevel's 1963 Children's Crusade against segregation or a reference to the Mexican National Crusade Against Hunger might be acceptable). Try to avoid words such as "sect" or "cult," because those words may be interpreted as demeaning to members of the groups cited.

Material about religion should be as objective as possible. Do not treat religion as a source of humor. Any focus on religion is likely to cause fairness problems if there is any plausible interpretation in which the material could be considered disparaging or negative. Furthermore, fairness problems are also likely if there is any plausible interpretation in which the material could be seen as positive or proselytizing. Be factually correct and neutral in any mention of religion, agnosticism, or atheism. Unless it is construct relevant, do not interpret one religion from the point of view of a different religion.

In tests made for a country that has an official religion, if the client requests religious material, it is acceptable to meet the request of the client as long as the material does not disparage other religions.

***Role Playing.*** Some constructed-response items ask test takers to assume a particular role and to respond from the perspective of a person in that role. Avoid construct-irrelevant roles that would cause test takers emotional distress. For example, do not ask test takers to assume the role of an enslaved person, a slaveholder, an inmate or guard at a concentration camp, a fired employee, an undocumented immigrant, or the like unless it is important for valid measurement. Do not ask test takers to take on construct-irrelevant roles that might be counter to their strongly held beliefs.

***Sexual Behavior.*** Avoid explicit descriptions of human sexual acts unless important for validity, such as in tests for medical personnel. Avoid double entendres and sexual innuendo unless important for validity, such as in literature tests for relatively mature test takers, and beware of inadvertent double entendres, especially in K–12 materials.

***Slavery.*** Avoid materials about slavery unless it is important for valid measurement, as in a history test. A brief mention of slavery in a passage used to measure a skill such as reading comprehension may be acceptable if it is clear that the passage is about something else. For example, a passage about the life and work of Mary McLeod Bethune might mention that her parents had been enslaved people.

Though "slave" is still an acceptable term, "enslaved person" is preferred (though note that "enslavement" is not an acceptable term for the general term "slavery"). "Slaveholder" is preferred to "slave owner." Authentic materials that use the terms "slave" and "slave owner" may be acceptable. Do not use materials with derogatory terms for enslaved people unless the materials are very important for validity and a more appropriate substitute is not available.

***Stereotypes.*** Avoid stereotypes (both negative and positive) in language and images unless they are important for valid measurement. Avoid using construct-irrelevant phrases that encapsulate stereotypes, such as "Dutch uncle," "Indian giver," "women's work," or "man-sized job." Avoid using words such as "surprisingly" when the surprise is caused by a person's behavior that is contrary to a stereotype. For example, avoid such sentences as "Surprisingly, a girl won first prize in the science fair."

Do not imply that all members of a group share the same attitudes or beliefs unless the group was assembled on the basis of those attitudes or beliefs. Avoid construct-irrelevant stereotypes in tests as sources of answer choices. Test takers who select an answer believe it is correct, so their belief in the legitimacy of a stereotype may be reinforced.

The terms "stereotypical" and "traditional" overlap in meaning but are not synonymous. Be careful when depicting an individual engaged in a traditional activity (such as a woman cooking). This does not necessarily constitute stereotyping as long as the test (or the item bank) as a whole does not depict members of a group engaged exclusively in traditional activities. If some group members are shown in traditional roles, other members of the group should be shown in nontraditional roles. A one-to-one balance is not necessary. To avoid reinforcing stereotypes, however, traditional activities should not greatly predominate.

In some rare cases, the need for valid measurement may acceptably reinforce a stereotype. For example, a test designed to certify nursing home assistants may find it necessary to depict most of the older residents as infirm and in need of help with the activities of daily life.

***Unstated Assumptions.*** Avoid material based on underlying assumptions that are false or that would be inappropriate if the assumptions had been stated. For example, do not use material that assumes all children live in houses with backyards, have access to local parks or swimming pools, or live with two parents. Do not use material that assumes all people over the age of 65 are retired and no longer have to work for a living.

As an example of inappropriate assumptions, consider the sentence “All social workers should learn Spanish.” The sentence is based on the unstated assumption that no social workers are native speakers of Spanish. There are additional unstated assumptions that speakers of Spanish have an inordinate need for the services of social workers, and that speakers of other languages have no need for the services of social workers who speak their languages.

Be careful using the word “we” unless the people included in the term are specified. The use of an undefined “we” implies an underlying assumption of unity that is often counter to reality and may make some test takers feel excluded. The people included in the term should be specified unless the use of an unspecified “we” is a common usage in the subject matter of the assessment.

***Violence and Suffering.*** Do not focus on violent actions, on violent crimes, on the detailed effects of violence, or on suffering unless such references are important for valid measurement. Violence and suffering are too widespread in art, biology, history, literature, and most aspects of human and animal life to exclude them completely from all material. For example, it is acceptable to discuss the food chain, even though it involves animals eating other animals. Do not, however, dwell unnecessarily on the gruesome or shocking aspects of violence and suffering.

***Visual Material.*** Do not use visual material (e.g., drawings, paintings, photographs, charts, graphs, diagrams, maps, videos) without a clear purpose for doing so. Unnecessary visual material can add to the cognitive load of an item, distract test takers from important information in the text, and make items less accessible for test takers with visual impairments. If visual material is used solely to make an item more engaging, weigh the increase in engagement against the need and ability to provide that same information in an alternate format (whether using descriptive text, tactile graphics, etc.).

When selecting visual materials, consider whether describing the image for people who are blind will result in excessive cognitive load and whether the graphic will be amenable to the creation of tactile graphics (e.g., raised-line drawings). Do not use variations in color or subtle differences in shading or pattern alone to indicate important distinctions, since this can be problematic for test takers with visual impairments.

Unless it is important for valid measurement, avoid visual material that depicts content out of compliance with the guidelines in this document. For example, the guideline about avoiding construct-irrelevant material that focuses on any religion applies to images of religious symbols.

Use images when they are construct relevant, but to improve accessibility and reduce unnecessary cognitive load, use the simplest images that are consistent with valid measurement and the need for authenticity. Avoid unnecessary visual clutter whenever possible. Because drawings can be simplified to contain only essential elements, they may be preferable to photographs when the realism and details of photographic images are not construct-relevant.

Scrutinize the background of visual material as well as the foreground when checking for fairness problems. Magnify the image as necessary (consider enlarging to 400%) to ensure that the entire image has been carefully inspected. Check reflections in windows, puddles, and so forth.

The clothing, facial expressions, gestures, and stances of any people in the image should be appropriate for the situation depicted and should not be likely to cause offense.

Avoid construct-irrelevant images of objects or actions that are controversial or offensive (e.g., a burning cross, a Confederate Battle flag, the Nazi salute, a swastika) or that may be mistaken for what are controversial or offensive images (e.g., the Buddhist swastika symbol). Refer to the section of the *GDFTC* titled "Cryptic References" for more details.

Avoid inappropriate gestures or hand signs that indicate obscenities, gang affiliation, anti-Semitism, White supremacist ideology, or the like (e.g., a middle finger raised is a common obscenity; a forefinger touching the thumb about halfway down, with the other three fingers spread, indicates "WP" for "White Power"; extending one finger on one hand and two fingers on the other hand indicates the letters "AB" for "Aryan Brotherhood"). Because there are many hand signs and because they are constantly changing, it is best to avoid construct-irrelevant images in which people are holding their hands or fingers in unnatural configurations. Note also that many seemingly innocuous images (e.g., eggplant, peach) may have sexual meanings in the world of sexting emojis.

If the images contain any text or numbers (e.g., graffiti,[8] banners, signs, posters, words on clothing or footwear, tattoos, etc.), make sure the content complies with the guidelines (e.g., no obscenities, no offensive or inflammatory statements, no brand names or logos unless construct relevant). Obtain translations of any text in a language you do not understand so that it can be evaluated. If you cannot obtain a translation, delete or obscure the text. If the answers to items depend on understanding the text or numbers in an image, the text or numbers should comply with the next section of the *GDFTC*, "Construct-Irrelevant Physical Barriers."

# 8.0 Construct-Irrelevant Physical Barriers

## 8.1 Requirements

ETS must meet the requirements for accessibility established in laws (e.g., the Americans with Disabilities Act and Section 508 of the Rehabilitation Act). Furthermore, ETS is committed to meeting the requirements for accessibility established in certain international standards (e.g., the Web Content Accessibility Guidelines, better known as WCAG[9]) for making information

[8] It is safest to exclude graffiti from images in K–12 tests unless the graffiti is construct-relevant.

[9] As of January 2022, the current official version is WCAG 2.1, though a working draft of WCAG 2.2 is available, the final draft is not scheduled to be released until June 2022. Be sure to use the most recent available version.

accessible on computers or other digital devices.[10] The goal is to provide the best measure of the tested construct for all test takers, offer an equitable test experience regardless of individual needs, and minimize the need for specialized accommodations.

Computer-delivered tests, related materials, and communications must be digitally accessible and compatible with assistive technologies.[11] Follow best practices for universal and inclusive design in the creation of tests and test products. Proper authoring of test items enables access with technologies or delivery modes such as audio, refreshable braille, and enlarged font. Paper-delivered assessments must be amenable to the creation of alternate formats.

### 8.2 Types of Physical Barriers

Construct-irrelevant physical barriers to success occur when aspects of tests that are not important for validity interfere with a test taker's ability to attend to, see, hear, or otherwise access the items or stimuli and/or to enter a response to the item. (This can be true as well for those receiving communications.) For example, test takers who are visually impaired may have trouble perceiving a diagram, even if they have the KSAs that are supposed to be tested by the item that is based on the diagram. Test takers with motor impairments may be unable to use an answer sheet or manipulate the input mechanism of a particular digital device, even if they have the KSAs measured by an item.

***Essential Aspects.*** Some aspects of tests are important or essential for validity and no acceptable substitute exists. "For example, it is reasonable to use a vision test as a requirement for a driver's license, even though the test is a physical barrier for aspiring drivers with poor vision. If no useful substitute is readily apparent, request an accessibility consultation from ETS's accessibility experts prior to finalizing that aspect of the test design.

***Helpful Aspects.*** Some aspects of various tests are helpful for measuring the intended construct, although supplementary content such as descriptive text might be needed to ensure meaningful access for individuals with disabilities. Those helpful aspects may be retained if mechanisms are provided to allow people with disabilities to respond appropriately to the item or task type. Items must be accessible to all test takers as is, or with one or more of the following:

- Universal tools that are available to all test takers as they choose. These tools may include such aids as a calculator, an English glossary, a highlighter, and magnification.

[10] These and other aspects of accessibility are explained in documents available to ETS staff. For further information, contact ACIS@ets.org.

[11] Assessments given online outside of testing centers must also take into account accessibility for students (or schools) who don't have access to computer or internet technology.

- Designated supports that are available to test takers as test accommodations. These tools may include such aids as a talking calculator, closed-captioning, adjustable colors, assistive technology, and special lighting.
- Changes to a test or its administration to make the test accessible for a person with a disability for whom the need is documented by an Individualized Education Plan (IEP), a Section 504 plan, or other documentation. Accommodations may include extra time, American Sign Language, a live reader, a scribe, paper large print, or paper braille.

***Unnecessary Aspects.*** Avoid unnecessary physical barriers in items and stimuli. Some physical barriers are simply not necessary. They are not important for valid measurement of the construct, nor are they even helpful in measuring the construct. Their removal or revision would not harm the quality of the item in any way. In many cases, removal of an unnecessary physical barrier results in an improvement in the quality of the item for all test takers. For example, a label for the lines in a graph may be necessary, but the use of a very small font for the label is an unnecessary physical barrier that could be revised with a resulting improvement in quality.

### 8.3 Examples of Physical Barriers

The following are examples of physical barriers in items or stimuli that may be unnecessarily difficult for test takers, particularly for people with certain disabilities. If these barriers, or others like them, are not important for validity, avoid them in items and stimuli:

- construct-irrelevant use of visually intensive tasks or tasks that require visually based mental manipulation of an object
- construct-irrelevant charts, maps, graphs, and other visual stimuli
- construct-irrelevant drawings of three-dimensional solids when a two-dimensional rendering would suffice, such as adding a meaningless third dimension to the bars in a bar graph
- construct-irrelevant measurement of spatial skills (visualizing how objects or parts of objects relate to each other in space)
- decorative rather than informative illustrations or parts of illustrations, such as decorative borders around images
- visual stimuli (e.g., charts, diagrams, graphs, maps) that lack sufficient color contrast or are more complex, cluttered, or crowded than necessary
- visual stimuli in the middle of paragraphs
- images of text rather than text itself (which creates a violation of the WCAG standards) unless essential to the task being performed

- visual stimuli as response options when the item could be revised to measure the same point equally well without them. Visual response options may be helpful, and therefore possibly acceptable, when used to reduce the reading load of an item; however, consideration must be given to the memory load that associated descriptive text would create.
- shading or color used alone to mark important differences in a visual stimulus
- lines of text that are vertical, slanted, curved, or anything other than horizontal
- text that does not contrast sharply with the background
- fonts that are hard to read and fonts for which it is impossible or difficult to distinguish among lowercase “l,” uppercase “I,” and the number “1,” if those distinctions are consequential
- letters that look alike (e.g., O, Q) or sound alike (e.g., s, x) used as labels for different things in the same item or stimulus
- numbers 1–10 and letters A–J used as labels for different things in the same item or stimulus, because the same symbols are used for those numbers and letters in braille and relevant braille symbol indicators might be overlooked
- special symbols or non-English alphabets, unless that is standard notation in the tested subject, such as Σ in statistical notation
- uppercase and lowercase versions of the same letter used to identify different things in the same item or stimulus, unless that is standard notation in the tested subject
- Roman numerals unless they are construct relevant. Screen readers do not reliably distinguish between Roman numerals and other groups of letters.
- the letter “A” as a variable in a math problem, because it is often voiced as “uh”
- long strings of italics or all capital letters and a mix of upper- and lowercase letters
- abbreviations for units of measurement in answer box labels (instead, use “inches” rather than “in” and “liters” rather than “L”)
- within certain math and science contexts, dashes in ranges of numbers, e.g., 9–27. Instead, use the word “through” in ranges (e.g., 9 through 27). In those same contexts, do not use the word “to” in a range of numbers, because it is easily confused with the number “two” when read aloud.
- centered text, especially when it may wrap onto the next line. Whenever possible, use left-justified text.

- the presentation of information in a table unless the use of a table has advantages over other ways of presenting the information. If tables are used, make them as simple as is consistent with valid measurement.

Some of the preceding examples may be acceptable if they are important for valid measurement or required for the authenticity of stimuli.

In addition, ensure that audio presentations are clear enough that the quality of the audio does not serve as a source of construct-irrelevant difficulty. Similarly, text and images displayed on a computer screen should be clear enough that the quality of the display does not serve as a source of construct-irrelevant difficulty.

Reduce the need to scroll to access parts of stimulus material or items to the extent possible, unless the ability to scroll is construct-relevant. If scrolling is required, however, make clear to the test taker that scrolling is necessary and provide instructions for how to do it.

Do not assume that all test takers will use a mouse or a keyboard to respond to items delivered on a digital device. Avoid using words that apply only to mouse users, such as “click on.” Instead, use a more general word, such as “select.” Use the word “enter” rather than “type” to accommodate various digital and assistive devices.

Because items may be delivered on multiple devices or with the use of different assistive technologies, the parts of the item may not maintain their intended spatial relationship for all test takers. Therefore, avoid referring to parts of an item as being above or below, or to the left or right of, other parts of the item. Instead, use general references such as “preceding” or “following.”

# 9.0 Appropriate Terminology for Groups

Language changes over time, and group preferences for group names change as well. As the changes occur, there is a transition period in which some group members prefer the older terminology and other group members prefer the newer terminology. Because one purpose of these guidelines is to avoid offending test takers, we have adopted a conservative stance toward words in transition.

If group identification is necessary, it is generally most appropriate to use the terminology that group members prefer. Unless very important for valid measurement, do not use names generally considered to be derogatory for groups, even if the names are used by some group members. Unless there is a reason not to do so, use the terminology adopted by the United States Census Bureau.

ETS recommends asking test takers to identify their race, ethnicity, or gender only if the data are to be used for an important purpose, such as studies of differential item functioning (DIF) or reporting average scores by group. ETS also recommends allowing test takers to select more than one response when asking test takers to identify their race or ethnicity. For gender, the traditional “male” and “female” options should, where possible, be augmented with other choices, such as “nonbinary,” “prefer to self-describe,” and “prefer not to answer.”

In general, use group names such as “Asian,” “Black,” “Hispanic,” and “White” as adjectives rather than as nouns. For example, “Hispanic people” is preferred to “Hispanics.” It is acceptable to use these terms as nouns sparingly after the adjectival form has been used.

Additionally, please note the following:

- Terms such as “African American” and “Native American” are not hyphenated, even when used as adjectives.
- The words “White,” “Black,” and “Indigenous” when referring to people are capitalized, but the word “people” in constructions such as “Indigenous people” is not.

The phrase “people of color” is not capitalized.

Discussions of appropriate terminology for various population groups follow. Some terms, such as “African American,” apply only to United States groups. For tests made for specific countries other than the United States, or for specific jurisdictions within the United States, determine the client’s preferences concerning terminology.

In authentic historical and literary material, some violations of the guidelines may be inevitable. Such material may be acceptable when it is construct relevant. Avoid materials with offensive and inflammatory terms, however, unless the materials are very important for valid measurement and more appropriate substitutes are not available.

### 9.1 People Who Are African American

The terms “Black” and “African American” are both acceptable, but not all Black people in the United States (e.g., some people from Caribbean countries) identify as African American. Note that African American is not hyphenated, even when used as an adjective. Note that “Black” should begin with an uppercase letter when referring to people. The terms “Afro-American,” “Negro,” and “colored” are not acceptable except when embedded in literary or historical contexts or in the names of organizations. The phrase “people of color” includes Black people as well as some other groups. Do not use “people of color” to refer to Black people in the absence of other groups. The relatively new term BIPOC (Black, Indigenous, and people of color) is also acceptable when that range of groups is being referenced. Because “Black” is used as a group identifier, try to avoid the use of “black” as a negative adjective, as in “black magic,” “black day,” or “black hearted.” Historical references such as “Black Friday” or “the Black Death” are acceptable when construct relevant.

### 9.2 People Who Are Asian American

The terms “Asian American,” “Pacific Island American,” “Asian/Pacific Island American,” and “Pacific Islander” should be used as appropriate. The term “Asian” includes people from many countries (e.g., Bangladesh, Cambodia, China, India, Japan, Korea, Laos, Pakistan, Thailand, Vietnam). Therefore, if possible, use specific terminology such as “Chinese American” or “Japanese American.” Do not use the word “Oriental” to describe people unless quoting historical or literary material or using the name of an organization.

### 9.3 People with Disabilities

To avoid giving the impression that people are defined by their disabilities, the generally preferred usage is to put the person first and the disabling condition after the noun (e.g., “a person who is blind”) in the first reference to a person or group. It is then acceptable to use disability-first terminology in later references. Some people with disabilities, however, prefer the disability-first terminology (e.g., “autistic person”). If you know which terminology is preferred by a person, use it in references to that.

Though the words and phrases may be impossible to avoid in literary or historic materials, try to minimize terms that have negative connotations or that reinforce negative judgments (e.g., “afflicted,” “confined,” “crippled,” “inflicted,” “pitiful,” “stricken,” “suffering from,” “victim,” or “unfortunate”). When possible, such terms should be replaced with others that are as objective as possible. For example, substitute “uses a wheelchair” for “confined to a wheelchair” or “wheelchair bound.” Similarly, try to avoid euphemistic or patronizing terms such as “special” or “physically challenged” as well as such words and phrases as “inspirational,” “courageous,” “overcoming a disability,” or “achieving success in spite of a disability.”

When possible, avoid the term “handicap” to refer to a disability. A disability may or may not result in a handicap. For example, a person who uses a wheelchair is handicapped by the steps to a building but not by a ramp or an elevator. Also try to avoid the term “handicap” to refer to an object that has been modified to make it accessible for people with disabilities. For example, refer to an “accessible toilet” rather than a “handicap toilet.”

Avoid implying that someone with a disability is sick unless that is the case. People with disabilities should not be called patients unless their relationship with a medical doctor is the topic. If a person is in treatment with a nonmedical professional (e.g., social worker, psychologist), “client” is the appropriate term.

Tests or other publications that deal specifically with teaching, diagnosing, or treating people with disabilities may require the use of certain terms with specialized meanings that might be inappropriate in general usage. The terms “normal” and “abnormal” referring to people are best limited to biological or medical contexts.

### 9.4 People Who Are Blind

It is preferable to put the person before the disability. The noun form “the blind” is best used only in the names of organizations or in literary or historical material. The phrase “visually impaired” is acceptable to cover different degrees of vision loss.

### 9.5 People with a Cognitive Disability

Preferable terms are “individuals with cognitive disabilities,” “developmentally delayed,” “developmentally disabled,” and “individuals with learning disabilities.” Use the term “Down syndrome” rather than “Down’s syndrome.” Do not use the obsolete terms “retarded” and “Mongoloid.”

### 9.6 People Who Are Deaf

The word “deaf” is acceptable as an adjective, but sometimes the terms “deaf” or “hard of hearing” may be used as a noun (e.g., School for the Deaf). The Deaf community and educators of individuals with hearing loss prefer “deaf and hard of hearing” to cover all gradations of hearing loss. References to the cultural and social community of Deaf people and to individuals who identify with that culture should be capitalized, but references to deafness as a physical phenomenon should be lowercase. Avoid the phrases “deaf and dumb, “deaf mute,” and “hearing impaired.”

### 9.7 People with a Motor Disability

The terms “motor disability” and “motor impairment” are both acceptable. The words “paraplegic,” and “quadriplegic” are acceptable as adjectives, not as nouns. The word “spastic” is unacceptable when used to describe a person.

### 9.8 People of Different Genders, Sexes, and Sexual Orientations

The general goal of this section of the *GDFTC* is to treat people equally regardless of their gender, sex, or sexual orientation.

“*Gender* refers to the attitudes, feelings, and behaviors that a given culture associates with a person’s biological sex” (APA, 2020, p. 138). Gender is a social identity and is not necessarily consistent with the sex assigned to a person at birth. The word “sex” refers to biological distinctions. “Sexual and romantic orientation” (referred to in this document as “sexual orientation") refers to the gender(s) or sex(es) of the people to whom a person is physically and/or romantically attracted and/or how a person feels attraction.

Do not use the phrase “sexual preference” for sexual orientation. Avoid the phrase “homosexual relationship,” and instead use “same-sex relationship.” Do not refer to heterosexual relationships as “normal” and other types of relationships as “abnormal.”

Do not assume that a pair or even a larger set of discrete categories necessarily includes the genders or sexes of all people. Avoid the phrases “the opposite sex,” “both sexes,” and “both genders" because they imply that only two possibilities exist. Do not assume that a group of adults consists only of men and women, or that a group of children consists only of boys and girls. Do not base the answer to an item on the unstated false assumption that the category "male" plus the category "female" always includes all people. Do not assume that a married couple necessarily consists of a man and a woman.

Avoid reliance on gender or sex as a distinguishing feature among people in items unless doing so is important for valid measurement, or unless other means of distinguishing among people increase the cognitive load and make the item more difficult for test takers to understand.

For a specific individual, use the term for sexual orientation selected by the individual, if it is known.

The adjective "gay" can be used to include all genders, or it can be used to include only men. Do not use "gay" as a noun. "Lesbian," however, may be used as an adjective or as a noun.

Avoid using the term "homosexual" outside of a scientific, literary, or historical context.

The term "queer" is gaining acceptance. It is still considered derogatory by some, however, except in reference to the academic fields of queer theory and queer studies in institutions that use those terms. Therefore, check the acceptability of the use of "queer" as a term for sexual orientation in the testing program in which the material will be used.

In general, when known, use the terminology preferred by the group.

"Trans" and "transgender" are acceptable as adjectives but not as nouns. These adjectives refer solely to a person's gender identity being inconsistent with that which is assigned at birth, not to an individual's sexual orientation. "Transgendered" is not acceptable. "Transsexual" is antiquated and considered offensive to many and is similarly not acceptable.

A common initialism for "lesbian, gay, bisexual, and transgender" is "LGBT." "LGBTQ" is used to add "queer" or "questioning," and "LGBTQIA" is used to add "intersex" and "asexual" or "allied" (however, some people may be puzzled by the "IA" in "LGBTQIA"). People often add a "+" to the abbreviation to stand for other groups (e.g., "LGBTQ+"). While "LGBT" is still acceptable, some people consider it to be outmoded. "LGBTQ+" appears to be the most common initialism at this time and is widely understood; it is therefore the preferred initialism. On balance, "LGBTQ" and "LGBTQ+" seem most appropriate, but the others may be used. In any case, it is best to define the initialism the first time you use it and to be sure it is representative of the groups about which you are writing.

When possible, people of all genders, sexes, and sexual orientations should be referred to in parallel terms. Do not, for example, refer to people of one gender by title and last name or by first and last name while people of another gender are referred to by first name only. Gratuitous, construct-irrelevant references to appearance or attractiveness of any people are not acceptable except in literary or historical material.

It is generally not appropriate to speak of human beings using “male” and “female” as nouns. Using the words as adjectives is acceptable, but not preferred.

Except in literary or historical material, people who identify as women and who are eighteen or older should be referred to as “women,” not “girls.” People who identify as men and who are eighteen or older should be referred to as “men,” not “boys.”

The term “ladies” should be used for women only when men are being referred to as “gentlemen.” Similarly, women should be referred to as wives, mothers, sisters, or daughters only when men are referred to as husbands, fathers, brothers, or sons.

“Ms.” is the preferred title for women, but “Mrs.” is acceptable in the combination “Mr. and Mrs.,” in historical and literary material, or if the person is known to prefer it. “Mx.” (pronounced “mix”), which is often used by nonbinary people, is gaining use as a gender-neutral title but may not be widely understood yet. Check its acceptability with the testing program in which the material will be used, but always use “Mx.” if the person is known to prefer it.

Using “he” or “man” to refer to all people is not appropriate unless the words are included in historical or literary material. Minimize the use of words that suggest that all members of a profession or all people serving in a role are male (e.g., use “police officer” rather than “policeman,” “human beings” rather than “mankind,” “supervisor” rather than “foreman”).

Before about 1970, it was generally considered correct to refer to all human beings as “man” and to use words such as “chairman,” “mankind,” and “manpower” based on that convention. Therefore, it is very difficult to find authentic materials written before then that comply with these guidelines. If it is necessary to use older literary or historical materials, some violations of the guidelines may be inevitable, but try to select materials that minimize the violations. Where useful, consider including a footnote explaining, for instance, the previous use of such terms that are now considered inappropriate.

Because generic terms such as “doctor,” “nurse,” “poet,” and “scientist” include all people in the occupation, modified titles such as “poetess,” “woman doctor,” or “male nurse” are not appropriate except in historical or literary material. Do not use expressions such as “the soldiers and their wives” that assume only people of a particular gender fill certain roles unless such is the case.

Do not couple generic role words with gender-specific pronouns unless a particular person is being referenced. Do not, for example, use terminology that assumes that all kindergarten teachers or food shoppers are women or that all college professors or car shoppers are men. Try to avoid materials that refer to objects (e.g., vehicles) using gender-specific pronouns except in historical or literary material.

If the sex or gender of a subject is not specified, avoid “he or she” or “his or hers” as pronouns. Alternating generic “he” and generic “she” is not appropriate, because neither word should be used to refer to all people. Avoid the constructions “he/she” and “(s)he.”

For a particular person, use the pronoun used by the person, if it is known. If the selected pronoun is not widely understood (e.g., “em,” “ze,” “hir”), it may be necessary to explain its meaning for test takers.

To avoid gender-specific pronouns, use plural constructions or constructions that avoid any pronoun. For example, instead of “Every test taker should sign his or her answer sheet,” use “Test takers should sign their answer sheets,” “Sign the answer sheet,” or “Sign your answer sheet.”

Also, many authorities on writing style (e.g., APA, 2020; AP, 2019; University of Chicago Press, 2017) now consider it correct to use singular “they,” “their,” “them,” “themselves” or “themself” to avoid using gender-specific pronouns (e.g., “Every test taker should sign their answer sheet”). Because some people still object to that usage, however, check its acceptability in the testing program in which the material will be used. Test takers should not be penalized for using singular “they,” even in formal writing.

When singular “they” is used, it takes a plural verb, just as singular “you” does. Avoid constructions in which the antecedent of singular “they” is unclear because a plural noun is a plausible antecedent.

### 9.9 People Who Are Hispanic American

The terms “Latino American” (for men and mixed-gender groups), “Latina American” (for women), and “Hispanic American” (for all genders) are acceptable and may be used as appropriate.

“Latinx” has been accepted by some as a gender-neutral term, but it has been rejected by others as artificial and is not widely used by the members of the group to which it refers. “Chicano” and “Chicana” are accepted by some as terms for Mexican Americans, but the terms have been rejected as derogatory by others. Acceptance tends to vary by region. Check with the testing program to determine whether or not the terms are acceptable in that program.

Where possible, use a specific group name such as “Cuban American,” “Dominican American,” or “Mexican American” as appropriate.

### 9.10 People Who Migrate to the United States

“Immigrant” and “migrant” are acceptable terms. For people who enter the United States without legal permission, use “undocumented immigrant” rather than “illegal alien.” Do not use “alien” as a noun to refer to an immigrant. Do not use “illegal” as a noun to refer to an undocumented immigrant.

### 9.11 People Who Are Members of One or More than One Racial/Ethnic Group

Members of what are generally called minority groups are becoming the majority in many locations in the United States and are the majority in many other countries. Therefore, although the terms are still acceptable, try to reduce the use of "minority" and "majority" to refer to groups of people. Depending on the context, consider using "underrepresented" or "groups whose members have historically been discriminated against" or "historically marginalized" for the former and "dominant culture" for the latter.

The terms "biracial" and "multiracial," as appropriate, are acceptable for people who identify themselves as belonging to more than one race or ethnicity. The term "people of color" (as well as the more recent term BIPOC [Black, Indigenous, and people of color] is acceptable for biracial and multiracial people or as a term for a collective group of non-White people (a mixed group of people who are African American, Asian American, Hispanic American, or Native American). "Colored people" is not acceptable except in historical or literary material or in the name of an organization.

### 9.12 People Who Are Native American

The terms "American Indian," "Native American," and "Indigenous people" are acceptable.[12] Whenever possible, it is best to refer to a people by the specific group names they use for themselves. However, that name may not be commonly known, and it may be necessary to clarify the term the first time it is used, as in the following example. "The Diné are still known to many other peoples as the Navajo." Many Native Americans prefer the words "nation" or "people" to "tribe." The words "squaw" to refer to a Native American woman and "buck" or "brave" to refer to a Native American man are not acceptable except in construct-relevant historical or literary material.

Avoid using the term "Eskimo" for people who are more acceptably called Alaskan Natives. More specific terminology, such as "Aleut," "Inuit," or "Yupik," may be used as appropriate. Indigenous people in Canada are often referred to as members of the First Nations.

Clients may differ in their requirements regarding the appropriate terminology to be used regarding people who are Native American. Check with the responsible assessment director for the fairness requirements of the client.

### 9.13 People Who Are Nonnative Speakers of English

There are several acceptable terms for nonnative speakers of English, but the terms differ in meaning and should be used appropriately. "Nonnative speaker" is the most general term. "English-language learner (ELL)" and "English learner (EL)" are the preferred terms for K–12 students who are not yet fully competent in English; however, NAEP has decided to use "English

[12] Respondents often misunderstand "Native American" as an option in questionnaire items to mean "a person born in America." It is safer to use "American Indian" or the combination "American Indian or Native American" as an option in questionnaire items.

learner (EL)," not "English-language learner (ELL)." The term "English as a second language (ESL)" applies to people who are learning English in an English-speaking environment, whereas "English as a foreign language (EFL)" applies to people who are learning English in a non-English-speaking environment. "Limited English proficient (LEP)" is a term generally limited to legislation.

Use "ESL," "EFL," and "LEP" as adjectives, not as nouns; for example, use "She is an ESL student," not "She is an ESL." It is preferable to use "ELL" or "EL" as an adjective to put the emphasis on the person rather than on the person's lack of English proficiency. For any of these abbreviations, the first appearance in text, whether as a noun or an adjective, should be accompanied by the spelled-out term in parenthesis. All instances thereafter can be abbreviated, even when the abbreviation is used to refer to people. However, do not use ESL, EFL, and LEP as nouns to refer to people.

### 9.14 People Who Are Older

It is best to refer to older people by specific ages or age ranges, such as "people age 65 and above." It is also acceptable to use the term "older people." Avoid using "aged" or "elderly" as a noun. Minimize the use of euphemisms such as "senior citizens" or "seniors." Avoid the word "senile" for a person with dementia. Avoid such disparaging qualifiers as "geriatric," "feeble," "grumpy," "decrepit," "doddering," and the like. Tests in certain content areas such as geriatrics may use terms such as "old-old" or "oldest-old" that are not appropriate in general usage.

### 9.15 People Who Are White

The terms "White" and "European American" are both acceptable. The term "European American" is preferred by some people because of its parallelism to "African American," "Asian American," "Native American," and so forth. "Caucasian" may be used in existing materials but is no longer a preferred term. Note that "White" should begin with an uppercase letter when referring to people (including in the construction "non-White") and should be used as an adjective. Capitalize "Indigenous" as in "Indigenous people." Note that we also capitalize as following: "non-White." "Anglo American" is ambiguous because it can refer to a person from England or to a White, non-Hispanic American. Use the word only when the meaning is clear from the context in which it is used. Because "White" is used as a group identifier, try to avoid the use of "white" as a positive adjective as in "white knight." Neutral uses of "white" as in "White House," "white collar," and "white water," are acceptable.

## 10.0 Representation of Diversity

If a test mentions or shows people, test takers should not be made to feel alienated from the test because members of their group are not included. Therefore, the ideal test (or item bank) would include members of the various groups in the test-taking population. While it is not feasible to include members of every relevant group in a test, strive to represent diversity in tests that mention people or show people in images.[13]

In addition, strive to represent diversity in the authors and reviewers of materials. If material is about a group of people, it is preferable, but not required, that the author be a member of the represented group.

### 10.1 Application

Follow the guidelines below for tests designed for use primarily in the United States. (Such tests may be administered worldwide.) The diversity reflected in tests made for a specific country other than the United States should be appropriate for the country for which the test is designed. Consult the client or the responsible assessment lead to determine the characteristics of the test-taking population to be reflected in a test made for a country other than the United States. Also, consult the client or responsible assessment lead to determine the diversity to be reflected in a test made specifically for a jurisdiction within the United States, such as a consortium of states, a state, a city, or a school district.

### 10.2 People with Disabilities

Roughly 26 percent (or 1 in 4) of the people in the United States have a disability of some type. If suitable for the subject matter, try to have about 10 to 15 percent of the items that depict people include people with (visible or invisible) disabilities. For example, people can be depicted with crutches, glasses, guide dogs, hearing aids, support canes, walkers, wheelchairs, white canes, and so forth, or they can be described, for example, as being bipolar or having attention-deficit/hyperactivity disorder (ADHD).

Be careful not to reinforce stereotypes when showing people with disabilities. For example, a picture of a person in a wheelchair in a work setting may be appropriate. However, showing a person who is in a wheelchair being pushed by someone else in a construct-irrelevant situation could reinforce the stereotype that people with disabilities are dependent and need help. Ideally, the focus will be on the person and the disability will be incidental rather than the focus of the image or text.

### 10.3 Gender Balance

In tests that predominantly measure skills rather than specific subject matter, women, men, and nonbinary people should be represented in ways that match the general population. In addition to roughly balancing numbers of people, the status of the people shown should be

---

[13] ETS has recently formed a special diversity panel, which is a good source for discussions around how best to represent people in images. Further information can be found by contacting ETSDiversityPanel@ets.org.

reasonably equivalent. A mention of a specific well-known man such as Albert Einstein in one item is not balanced by a mention of a generic female name in another item.

The gender balance of tests that predominantly measure content should be appropriate to the subject matter. For example, most of the people mentioned or shown in a test of military history would be men. The gender balance of occupational tests should be sure to include some diversity, but the balance should very roughly approximate the gender distribution of members of the occupation.

### 10.4 People Who Are LGBTQ+

Avoid negative or derogatory references to people who are LGBTQ+ on the basis of their orientation. Avoid the implication that LGBTQ+ people are less valuable members of society than are other people.

Identify the sexual orientation and/or gender identity of people in tests for purposes of representing diversity as appropriate.

If there is insufficient context in an item to indicate group membership in other ways, representation may be accomplished by using the names of well-known people who are LGBTQ+ for purposes of representing diversity if the client approves. For example, you could use a sentence about James Baldwin to test a point of grammar.

You may also refer to same-sex couples (e.g., a woman and her wife) if straight couples are similarly identified. Additionally, where visual representations of people such as avatars are used, try to include gender-nonconforming and/or gender-neutral people. If offering a choice of avatars to represent the test taker, include a gender-neutral option.

### 10.5 Racial and Ethnic Balance

Because about one-third of the people in the United States are members of underrepresented groups, try to represent people from those groups in the United States or people from the countries of origin of those groups in about one-third of the items that mention or show people. For example, include African American people or African people, Asian American people or Asian people, and Latino people from the United States or from Latin America. Also include indigenous groups from the United States or from other countries. Items that include other groups that could be considered minorities, such as Americans of Middle Eastern origin or Middle Eastern people, may be counted among the items that represent diversity.

If there is insufficient context in an item to indicate group membership in other ways, representation may be accomplished by using well-known real people in various groups or by using generic names commonly associated with various groups.

In tests that predominantly measure specific subject matter, try to meet the preceding representational goals to the extent suitable for the subject matter, unless different requirements are stated in the test specifications. If the names of people appearing in tests are part of the subject matter (e.g., Avogadro's number, Heimlich maneuver, Monroe Doctrine), the items are not counted as mentioning people for the purpose of calculating the number of items in which diversity should be represented.

### 10.6 Societal Roles

If it is possible to do so in the materials for a test, demonstrate that people in different groups are found in a wide range of societal roles and contexts. It is best to avoid language and images that suggest that all members of any single group are people in higher-status positions or lower-status positions. For example, do not portray all of the executives as men and all of the support staff as women.

## 11.0 Fairness of Artificial Intelligence Algorithms

### 11.1 Background

At ETS, the rapidly increasing use of artificial intelligence (AI) algorithms in test development, constructed-response scoring, education, and related fields has raised new concerns about fairness.[14] Just as the fairness of a test depends on its validity for different groups of people, the fairness of an AI algorithm depends on the extent to which it appropriately meets its purpose for the different groups of people in the population affected by its use. As with the results of tests, group differences in the results of using an AI algorithm are fair if the differences in the results are based on real and relevant differences among the affected groups. Even though a computer algorithm appears to be objective, the use of AI is not necessarily fair unless appropriate precautions are taken.

AI systems use computers to perform tasks that previously required human judgment to perform (e.g., scoring essays, assembling tests, recommending books that match the interests and reading levels of children, writing test items, defining learning progressions). AI systems provide a substitute for human judgment by applying models to data, with the goal of matching a criterion consisting of human judgments. The data, for example, could be alphabetic and numeric characters, spaces, and punctuation marks constituting an essay or an item pool; audio signals constituting a spoken response; or pixels constituting a video.

The models applied to the data could consist of rules that were explicitly coded by a human being or of a set of statistical parameters and decision-making criteria that were generated by the algorithm itself in a form of machine learning. Machine-learning algorithms use repeated trial and error as input in order to generate the algorithm's output, which can be described as a

---

[14] The point at which a computer algorithm becomes "intelligent" enough to qualify as AI is open to debate. We use the term "AI" for algorithms that emulate human judgments. For example, an algorithm that scores a multiple-choice test is not considered AI, because a human scorer would not have to use judgment to match a list of the test taker's answer choices with a list of correct answer choices. An algorithm that scores an extended essay response, however, would be considered AI, because a human scorer would have to use judgment to determine the score.

set of model parameters used to match the criterion judgments. Parameters that improve the match when applied to the data are strengthened; parameters that worsen the match are dropped. Over repeated trials, the algorithm "learns" which parameters work best.

The models are limited to manipulating various weightings and combinations of variables that can be processed by a computer. Therefore, AI algorithms often use substitutes or proxies for actual variables of interest. For example, a computer cannot actually judge the creativity of an essay. A computer algorithm can, however, determine the computer-countable characteristics that differentiate between essays that have been given high creativity scores by human judges and essays that have been given low creativity scores. The algorithm may find that essays with high human-given creativity scores tend to use mixtures of long and short sentences, and essays with low human-given creativity scores tend to use sentences with only small differences in length. The algorithm could then use standard deviation[15] of sentence length as one of the proxies for creativity.

It is important to keep in mind that a proxy is associated with the relevant variable of interest, but the proxy is not the same as that variable. An algorithm based on standard deviation of sentence length may help match human judgments of creativity, but it is very clearly not the same as creativity. Therefore, proxies may be misleading, possibly resulting in an error that, in turn, might be a source of bias. Furthermore, if the people affected by an algorithm learn of the proxies, they may subvert the algorithm by changing their behavior relative to the proxies rather than to the actual variables of interest.

The use of some AI systems may not affect people directly, but there remains a concern for fairness because people are affected indirectly. For example, an AI system used to identify good sources for reading-comprehension passages may select passages that favor certain groups and disfavor other groups when the passages are used in tests. Therefore, fairness remains a concern whenever people are affected by the results of an AI system, whether directly or indirectly.

### 11.2 Bias

Because an AI system depends on the application of models to data, bias can be caused by inappropriate models, inappropriate data, or both. Some of the causes of bias include reliance on biased human judgments as a criterion, poorly sampled criterion data, and models based on inappropriate proxies.

***Biased Judgments.*** Most AI systems use the results of human judgment as a criterion to emulate. An example of such a criterion prevalent at ETS is the use of AI to score essays. A "training set" of human-scored essays is required. The automated scoring algorithm applies

[15] The standard deviation is a statistic that indicates how far apart the numbers in a distribution are. If the numbers are packed tightly together, the standard deviation is small. As the numbers become further apart, the standard deviation increases.

a model to the essays to match the human scores as closely as possible. Clearly, if the human-produced scores used as a criterion are biased, an algorithm built to match them will tend to produce similarly biased scores as well.[16]

***Poorly Sampled Data.*** Even if the judgments forming the criterion are fair, poorly sampled data forming the criterion sample may cause the AI algorithm to produce biased results. Consider, for example, an AI system used to counsel college students about occupations that would match their interests and abilities. One component of the algorithm compares the students' interests to the interests of criterion samples of members of various occupations. If almost all of the members of the criterion sample for an occupation were people with interests often ascribed to men, then the algorithm would not suggest that a student with interests often ascribed to women should consider the occupation, resulting in gender bias.

Poorly sampled data can also result in an algorithm that capitalizes on random characteristics that happen to be useful predictors in the criterion sample but are not replicated across samples. For example, in the training set of videos featuring images of cats and dogs for an image-recognition algorithm that feature images of cats and dogs, almost all of the cats may happen to be black and almost all of the dogs may happen to be brown. The system will "learn" to identify images of black, four-legged, furry things with tails as cats and will tend to misidentify brown cats and black dogs.

***Inappropriate Proxies.*** Because a machine-learning algorithm generates its own rules, it may capitalize on inappropriate variables that are correlated with ethnicity, gender, race, religion, and so forth, resulting in biased results. For example, the algorithm may "learn" that zip codes are associated with criterion judgments and use zip codes in its self-generated rules. Because of de facto segregated housing, zip codes serve as a proxy for race and will result in biased decisions based on race, even if the algorithm never used race as a variable.

### 11.3 Guidelines

Because fairness depends on the extent to which an AI system meets its goals for different groups of people, it is necessary to be clear about the goals of using the system. Therefore, stipulate the intended purposes of an AI algorithm, and identify the kinds and range of materials and the population of people for whom it is intended. Document the process by which the algorithm was developed. Describe the major decisions that were made and the qualifications of the people involved. Describe how fairness will be addressed in design, development, and use.

### 11.4 Consider Risks of Bias When Selecting AI Factors

When selecting the factors to be considered by AI, such as how the algorithm will evaluate input data, consider the risk of bias. To help mitigate such risks, document the factors and the

[16] For an extensive discussion of AI (and human) scoring, see the document "[Best Practices in Constructed-Response Scoring](#)."

criteria for choosing them. Some questions that might be relevant in this regard include the following:

- What factors should the algorithm consider?
- What initial weightings should be assigned to the chosen factors?
- What will ETS gain by developing the algorithm?
- How open will the design process be?
- Is the design team representative enough to capture and address the nuances of different cultural contexts? If not, what other steps can be taken to ensure sufficient representation?

To the extent possible, the algorithm should use only actual variables of interest rather than proxies for those variables. Provide a rationale to justify the use of any substitution or proxy for a relevant variable.

If human judgments are used as a criterion to be simulated by an AI algorithm, evaluate the judgments to help ensure that any group differences based on the human judgments are based on real and relevant differences among the groups.

### 11.5 Evaluate the Data Used to Train AI

Review the data that are used to train AI for accuracy. Ensure that there are sufficient data to accomplish the objective in a non-biased way. Specific questions might be relevant, depending on the use of the AI, and should be answered in order to evaluate whether training data are or are not likely to be biased. Questions may include the following:

- Where does/will the training data come from?
- Who is responsible for the collection and maintenance of the training data?
- Who does the training data cover? Does it reflect the intended population?

To guard against capitalization on random characteristics of the training sample for an AI algorithm, cross-validate[17] the algorithm by using a different sample.

### 11.6 Additional Evaluation and Documentation

Obtain and document evidence that the AI algorithm is meeting its intended purpose for the intended population. If use of the algorithm has direct or indirect consequences for people, assess the effects of using the algorithm on groups of interest, including people with disabilities and people who are not native speakers of English.

---

[17] To cross-validate is to evaluate the algorithm developed on one set of data through the use of an independent set of data.

Evaluate the extent to which group differences in results are based on real and relevant differences in the groups. If the intended use of an algorithm has unintended negative consequences for some group, review the evidence to determine whether or not the negative consequences follow from real and relevant group differences. Revise the algorithm to reduce inappropriate group differences. Document what has been done to address fairness and any ongoing fairness efforts, such as gathering data on the results of using the algorithm for different groups of people.

Periodically review the performance of active AI algorithms to verify that they continue to be appropriate. The appropriate time interval between reviews depends on judgments about the stability of the population, the stability of the algorithm, and the stability of the results of using the AI. For example, an algorithm that keeps on “learning” as it is continuously updated with new data should be reviewed much more frequently than an algorithm that remains stable. Select an appropriate frequency for review of the algorithm, and provide a rationale for the selected frequency.

Seek to develop and use what is commonly referred to as “explainable AI.” Explain how any conclusions are reached, and the basis for actions taken, at least to the extent of explaining what variable or variables drove the decisions that were determined through the use of AI. To the extent that explainable AI cannot be achieved, develop AI that is auditable so that the claims made on the basis of the tests can be evaluated.

State the limitations as well as the benefits of the algorithm. Warn intended users of potential misuses of the algorithm.

# 12.0 Additional Guidelines for Fairness of NAEP and K–12 Tests

These guidelines for NAEP and K–12 tests are in addition to, not a replacement for, the guidelines that apply to all ETS tests.

## 12.1 Requirements for NAEP

The following requirements are excerpted from the National Assessment Governing Board (NAGB) Policy Statement, NAEP Item Development and Review, adopted May 18, 2002. (As of February 2021, NAGB had not yet updated the statement. Check www.NAGB.gov for the latest version.)

***Secular, Neutral, Nonideological.*** Items shall be secular, neutral, and non-ideological. Neither NAEP nor its questions shall advocate a particular religious belief or political stance. Where appropriate, NAEP questions may deal with religious and political issues in a fair and objective way.

The following definitions shall apply to the review of all NAEP test questions, reading passages, and supplementary materials used in the assessment of various subject areas:

***Secular.*** NAEP questions will not contain language that advocates or opposes any particular religious views or beliefs, nor will items compare one religion unfavorably to another. However, items may contain references to religions, religious symbolism, or members of religious groups where appropriate.

Examples: The following phrases would be acceptable: “shaped like a Christmas tree,” “religious tolerance is one of the key aspects of a free society,” “Dr. Martin Luther King, Jr., was a Baptist minister,” and “Hinduism is the predominant religion in India.”

***Neutral and Nonideological.*** Items will not advocate for a particular political party or partisan issue, for any specific legislative or electoral result, or for a single perspective on a controversial issue. An item may ask students to explain both sides of a debate, or it may ask them to analyze an issue or to explain the arguments of proponents or opponents, without requiring students to endorse personally the position they are describing. Item writers should have the flexibility to develop questions that measure important knowledge and skills without requiring both pro and con responses to every item.

Examples: Students may be asked to compare and contrast positions on states’ rights, based on excerpts from speeches by X and Y; to analyze the themes of Franklin D. Roosevelt’s first and second inaugural addresses; to identify the purpose of the Monroe Doctrine; or to select a position on the issue of suburban growth and cite evidence to support this position. Or, students may be asked to provide arguments either for or against Woodrow Wilson’s decision to enter World War I. A NAEP question could ask students to summarize the dissenting opinion in a landmark Supreme Court case.

The criteria of neutral and nonideological also pertain to decisions about the pool of test questions in a subject area taken as a whole. The National Assessment Governing Board shall review the entire item pool for a subject area to ensure that it is balanced in terms of the perspectives and issues presented.

***Sensitive Topics.*** In addition to being secular, neutral, and nonideological, NAEP items should not discuss and must avoid asking students to reveal information about any of the following potentially sensitive topics:

- political affiliations or beliefs of students or family members
- mental or psychological problems of students or family members
- sexual behavior or attitudes
- illegal, anti-social, self-incriminating, or demeaning behavior
- critical appraisals of other individuals with whom there is a close family relationship, or a legally recognized privileged relationship, or analogous relationships, such as with a lawyer, physician, or clergy member

- religious practices, affiliations, or beliefs of students or family members
- income (other than required to determine eligibility for program or financial assistance)

NAEP repeats items for many years (some dating back to the 1970s) to allow the measurement of changes in the average knowledge and skills of the student population over time. These long-term-trend items were judged to be appropriate when first used, but they may not meet all current guidelines. As previously noted, however, "any material that is important for valid measurement—and for which a similarly important but more appropriate substitute is not available—may be acceptable for inclusion in a test, even if it would otherwise be out of compliance with the guidelines."

Because the older items are required to measure change over time, newer materials would not be valid for that purpose. Therefore, the older items used to measure trends over time are generally acceptable even if they do not comply with all current fairness guidelines.

In addition to the requirements set by the National Assessment Governing Board and the guidelines for all ETS tests, NAEP follows the guidelines for K–12 assessments described below.

### 12.2 K–12 Assessments

K–12 assessments include tests commissioned by consortiums of states, individual states, cities, or school districts for use in their classrooms from kindergarten through the end of high school. The fairness requirements for K–12 tests are often more rigorous and extensive than the fairness requirements for other tests. Many jurisdictions are extremely cautious about the content of K–12 tests because young children will be exposed to the material. Furthermore, various constituent groups within a jurisdiction may have very strong beliefs about acceptable test content, which are reflected in the jurisdiction's fairness guidelines.

The fairness guidelines that apply to all ETS tests are not repeated here, but those guidelines must be followed in addition to the jurisdiction's requirements.

The following guidelines have been compiled from the requirements of several state clients and may serve as an overview of the types of issues of particular concern for K–12 tests. Different K–12 clients, however, may have different fairness requirements. New clients may have additional requirements not listed here, and requirements of existing clients may change over time. Political sensitivities may change, the staffing of departments of education may change, and the membership of fairness review committees may change, leading to additional or revised fairness requirements. Therefore, check with the assessment director who is responsible for the current fairness requirements of the client. It is very helpful to document the fairness requirements of the client for use by item writers and reviewers.

In developing assessments for K–12 testing, it is important to avoid topics to which certain groups of students may be especially sensitive. Many topics are considered inappropriate for tests in certain jurisdictions, even though they may be discussed in classrooms.

***Emotionally Charged Topics.*** Unless they are important for validity, avoid discussions of topics that may be excessively emotionally charged for K–12 students, such as

- problems caused by uncommon physical or emotional attributes (e.g., anorexia, disfigurement, early or late physical development, obesity, small stature, stuttering);
- dissension among family members, between students and teachers, or between parents/guardians and teachers;
- serious illnesses or widespread infections (e.g., cancer, COVID-19, Ebola, herpes, tuberculosis) or death, particularly of children, siblings, or parents (it is acceptable to mention the death of historic figures, e.g., "President Kennedy died in 1963");
- natural disasters, such as earthquakes, hurricanes, tornadoes, floods, or forest fires, unless the disasters are treated as scientific subjects and there is little mention of the destruction caused and loss of life;
- segregated schools or neighborhoods, ghettos, slums;
- family situations that students may find upsetting (e.g., deportation of a parent or sibling, divorce, separation, eviction, homelessness, incarceration, layoff, job loss);
- human trafficking, forced labor, sexual exploitation;
- technology that results in loss of jobs;
- violence or conflict, including domestic violence, playground arguments, fights among students, bullying (including bullying through social media), cliques, and social ostracism;
- graphic violence in the animal kingdom, a focus on pests (e.g., rats, roaches, and lice), or a focus on the threatening aspects of creatures that may be frightening to children (e.g., poisonous snakes, spiders); and
- animals that may be sensitive topics for specific cultural groups (e.g., the owl for some Native American nations). Check with the client or the responsible assessment director to verify whether references to animals of any sort may be a sensitive topic.

***Individual and Group Names.*** Clients differ on how they want the names of people to be used in test materials. Some clients want to replace names with gender-neutral references such as "the student" or "the teacher." Other clients prefer to use names as an opportunity to represent diversity in test materials and encourage the use of names such as "Ms. Ramos," “Jazmin,” “Hiroshi,” “Raji,” and the like. Some clients prefer what are common, simple names in American English, such as “Mr. Smith,” “Ann,” and “Bob.”

Similarly, clients may differ on the preferred names for groups. For example, “Chicano” and “Chicana” may be acceptable in one jurisdiction but not in another. One jurisdiction may prefer “Black” and another may prefer “African American.” “Latinx” is not widely accepted at the K–12 level, but some clients may allow it. NAEP has decided to use EL (English learner) instead of ELL (English-language learner).

Ask the program assessment director to identify the program’s preferred terms for groups and names for individuals.

***Offensive Topics.*** Unless they are important for validity, avoid topics such as the following that may be offensive to various groups in a jurisdiction:

- drinking alcohol, smoking, vaping, chewing tobacco, using drugs (including prescription drugs in some jurisdictions)
- gambling. Some clients do not allow the use of playing cards or dice in stimuli. Even if references to these implements are allowed, do not assume that all students will be familiar with them.
- holidays and other occasions not observed by some students (e.g., birthday celebrations, Halloween, religious holidays, Valentine’s Day)
- social dancing, including school dances (such as proms); certain kinds of music (e.g., punk, rock and roll) and controversial lyrics; attending movies. These sensitivities vary greatly by client.
- references to a deity, including expressions like “thank God” and euphemisms such as “geez” or “gee whiz.” While it is appropriate to include literature and texts from many cultures, it is best to avoid stories about mythological gods or creation stories, unless the client requests them.
- extrasensory perception, UFOs, the occult, or the supernatural
- texts that are preachy or moralistic, because they may offend populations that do not hold the values espoused

***Controversial Topics.*** In addition to controversial topics discussed as best avoided for all ETS tests, there are many controversial topics that are best excluded from K–12 testing. Do not promote or defend personal or political values in K–12 test materials. Maintain a neutral stance on controversial issues unless the jurisdiction’s standards require stimuli that are designed to be persuasive or controversial. Some clients may want such passages or stimuli to be clearly labeled as persuasive or editorial text. Topics that are particularly troublesome in some jurisdictions include

- animals (keeping animals in zoos or theme parks, any possible implication of mistreatment of an animal);
- deforestation, environmental protection, global warming, human contribution to climate change;
- evolution (with associated topics of natural selection), fossils, geologic ages (e.g., millions of years ago), dinosaurs, and similarities between people and other primates, unless required by content standards;
- gun control;
- human trafficking;
- immigration, deportation, Immigration and Customs Enforcement (ICE), forced migration, treatment of immigrants;
- labor unions;
- meat consumption, raising animals for food, slaughtering animals;
- patriotism and the American dream, which are not always seen as positive concepts;
- police activity, law enforcement;
- prayer in school;
- racism, sexism, ageism, the suffering of individuals at the hands of a prejudiced society, a focus on individuals overcoming prejudice, or the specific results of discrimination;
- robots leading to loss of jobs;
- theme parks;
- vaccination; and
- welfare or food stamps

***Inappropriate Behavior.*** Do not use material that models or reinforces inappropriate student behaviors. Do not make such behaviors appear to be fun, attractive, rewarding, glamorous, sophisticated, or pleasurable. Such behaviors include

- trying to deceive teachers or other adults, lying, stealing, running away from home, or even considering those behaviors;
- bullying, cyberbullying, inappropriate use of the Internet;
- going without sleep, failing to attend school or do homework, or eating large quantities of junk foods;

- violating good safety practices (e.g., keeping dangerous animals, entering homes of unknown adults, using weapons or dangerous power tools; for younger children, activities such as baking, using scissors, and walking or biking to other than routine destinations without appropriate supervision);
- breaking laws or school rules;
- sexual activity, unless required by content standards;[18] and
- expressing or implying cynicism about charity, honesty, patriotism, reverence, or similar values esteemed by the community.

***Specific Content Areas.*** Material that is important for validity should be included in a test. Therefore, any topic that is required by a jurisdiction's content standards may be included in a test, even if it has been described as a topic best avoided.

The subject of battles and wars, for example, usually cannot be avoided in social studies tests at grade 5 and above. Slavery is a similar issue that may be appropriately addressed within certain content standards. A discussion or description of disease may be necessary in science or health assessment. If topics such as disasters, disease, slavery, terrorism, or war are required by the state's content standards, the topics should be presented in a manner sensitive to the feelings of students who may have strong emotions concerning those issues.

Some jurisdictions may require the use of certain genres (such as fables and myths) in reading assessments, or they may require the use of literature of historical or literary importance. Such older material may reinforce stereotypes, use nonparallel terms for different genders, use generic "he" to include all people, use words such as "fireman" that assume only men fill certain roles, model inappropriate behavior, or focus on conflict and the like. Material important for validity is acceptable, but strive to find the least problematic material that meets the jurisdiction's content requirements.

# 13.0 Conclusion

The *GDFTC* deals with only one aspect of the fairness of tests and communications—the appropriateness of the content and images. Fairness in assessment has many other aspects, however, including how test takers are treated, how items are made accessible for test takers with disabilities and for English learners, how tests are administered, how extended responses such as essays are scored, how items and tests are analyzed, how scores are made comparable, how scores are reported, and how the scores are used. All of these additional aspects of fairness are clearly beyond the scope of the *GDFTC*. Interested readers should consult, for example, AERA, et al. (2014); Camilli (2006), Dorans & Cook (2016); ETS (2010); ETS (2014);

[18] Note that sexual *orientation*, which is part of one's internal and individual sense of attraction, is distinct from sexual *activity*. Mentions of a person's sexual orientation should not be excluded on the grounds of "sexual activity" listed here.

Osterlind & Everson (2009); and Zwick (2002). In addition, a large number of reports of research done at ETS on fairness and other measurement-related topics are available for free at www.ets.org. Select the tab for Research and use the search function in the "Find a Publication" search field near the top of the page.

The task of developing guidelines for the fairness of tests, communications, and other products and services is never truly completed. What is considered fair changes over time, so some of these guidelines will eventually become obsolete and new guidelines will have to be added.

It is impossible to develop guidelines and examples for fairness that will cover every situation, and reasonable people can disagree about what fairness means. Therefore, judgment is required to interpret the guidelines for a specific product that has been created for a particular purpose and for a certain population. If appropriately interpreted and applied, however, the GDFTC will help ETS attain the goal of making tests and other products and services as fair as possible.

# 14.0 References


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

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th edition). Washington, DC: Author.

Associated Press. (2019). *The Associated Press Stylebook*. New York: Basic Books.

Camilli, G. (2006). Test fairness. In R. L. Brennan, (Ed.). *Educational measurement* (4th edition, pp. 221–256). Westport, CT: Praeger.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115–124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement*, 10, 237–255.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement*, 38, 369–382.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43 (6), 1241−1299.

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

Dorans, N., & Cook, L. (Eds.). (2016). *Fairness in educational assessment and measurement*. New York: Routledge.

ETS. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Hakkinen, M. (2015, June). Assistive technologies for computer-based assessments. *R&D Connections*, 24, 1–9.

Holland, P., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th edition, pp. 17–64). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43, 139–161.

Madiega, T. (2019). *EU guidelines on ethics in artificial intelligence: Context and implementation* (PE 640.163). European Parliamentary Research Service. Retrieved from https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd edition, pp. 13–103). New York, NY: Macmillan.

Mogilner, A., & Mogilner, T. (2006). *Children's writers' word book* (2nd ed.). Cincinnati, OH: Writer's Digest Books.

Organization for Economic Cooperation and Development (OECD). (2019). *Recommendation of the Council on Artificial Intelligence, section 1, article IV, 1.3*. Retrieved from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#_ga=2.100756866.2085259285.1559316172-1236900936.1559134188

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Thousand Oaks, CA: Sage.

Shepard, L. A. (1987). The case for bias in tests of achievement and scholastic aptitude. In S. Modgil & C. Modgil (Eds*.), Arthur Jensen: Consensus and controversy*. London, England: Falmer Press.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Birsner, E. P. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies*. Austin, TX: Steck-Vaughn Company.

Thorndike, R. L. (1971). Concepts of cultural fairness. *Journal of Educational Measurement*, 8, 63–70.

University of Chicago Press. (2017). *The Chicago manual of style*. Chicago: Author.

World Wide Web Consortium. (2018). *Web content accessibility guidelines 2.1* Retrieved from http://www.w3.org/TR/WCAG21/

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: Erlbaum.

Zieky, M. J. (2013). Fairness review in assessments. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 293–302). Washington, DC: American Psychological Association.

Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.). *Handbook of test development* (2nd edition, pp. 81–99). New York: Routledge.

Zwick, R. (2002). *Fair game?: The use of standardized admissions tests in higher education*. New York: Routledge Falmer.

# 15.0 Glossary

**Accommodation** – A change to a test, or to its administration, to assess the intended construct for a person with a disability. An accommodation does not change the intended construct. Refer to Construct, Disability. Compare Modification.

**Adaptation** – Changes to a test to make it suitable for a culture other than the one for which it was designed. Refer to Test.

**Administration Mode** – The method by which a test is presented to the test taker, including, for example, printed booklets, braille booklets, American Sign Language, computer display terminals, audio files, and videos. Refer to Test. Compare Response Mode.

**Algorithm** – A step-by-step procedure for accomplishing some activity. Often, algorithms are coded for implementation by computer.

**Alternate Form** – Different editions of the same test, written to assess the same skills and types of knowledge at the same level of difficulty, but with different tasks, questions, or problems. Refer to Task, Test.

**Ancillary Materials** – Descriptive booklets, score interpretation guides, administration manuals, registration forms, etc., that accompany a test. Refer to Test.

**Artificial Intelligence** – The use of computer algorithms to simulate human cognitive skills such as judgment, analysis, and problem solving. Refer to Algorithm.

**Assessment** – A synonym for test. In some usages, assessment refers to a broader process of evaluation and judgment. Refer to Test.

**Automated Item Selection** – A computer algorithm that follows a set of rules to assemble a test that meets specifications from a pool of items. Refer to Algorithm, Item, Specifications.

**Automated Scoring of Constructed Responses** – The use of a computer algorithm to generate a score for an essay or other constructed response. (The computer program that implements the algorithm is often referred to as a “scoring engine.”) Refer to Algorithm, Constructed-Response Item.

**Bias** – In general usage, unfairness. In technical usage, the tendency of an estimation procedure to produce estimates that deviate in a systematic way from the correct value. Refer to Fairness.

**Client** – An agency, association, organization, institution, individual, jurisdiction, or the like that commissions ETS to provide a product or service.

**Cognitive Test** – A test of knowledge or intellectual skill. Compare Noncognitive Test.

**Computer-Based Test** – Any test administered on a computer (also known as Computer-Based Assessment, or CBA).

**Construct** – The set of knowledge, skills, abilities, or other attributes a test is intended to measure, such as knowledge of American history, reading comprehension, study skills, writing ability, logical reasoning, honesty, calculus, intelligence, and so forth.

**Construct-Irrelevant Variance** – Differences among test takers' scores that are caused by factors other than differences in the knowledge, skills, abilities, or other attributes included in the construct that the test is intended to measure. Refer to Construct, Variance. Compare Construct-Relevant Variance.

**Construct-Relevant Variance** – Differences between test takers' scores that are caused by differences among test takers in the knowledge, skills, abilities, or other attributes included in the construct the test is intended to measure. Refer to Construct, Variance. Compare Construct-Irrelevant Variance.

**Constructed-Response Item** – An item in which answers are generated by the test taker rather than selected from a list of possible responses. Refer to Item. Compare Selected-Response Item.

**Context** – In an item or stimulus, the non-tested setting or situation in which some skill is to be applied. For example, an item about calculating the area of a rectangle might be set in the context of painting a wall. Refer to Item, Stimulus.

**Criterion** – That which is predicted by a test, such as college grade-point averages or job performance ratings. Refer to Test.

**Cross-Validate** – To repeat a study but with a different sample to determine whether the results are consistent. Refer to Sample.

**Customer** – A general term for those who sponsor, purchase, or use ETS products or services, including clients, institutional and individual score recipients, and test takers.

**Differential Item Functioning (DIF)** – Any tendency of a test item to be harder or easier for members of a particular group of test takers than for equally able members of another group. Generally, in measures of DIF, members of different groups are considered equally able on what a test is measuring if they receive the same total score on the test. In most DIF analyses, the groups of test takers are defined on the basis of gender, race, or ethnicity. Refer to Item, Test. Compare Impact.

**Disability** – A physical or mental impairment that substantially limits a major life activity. Individuals with disabilities may request accommodations or modifications in order to have meaningful and equitable access to a standardized test. Refer to Accommodation, Modification, Standardized Test.

**Discrimination** – The power of an item to differentiate among test takers at different levels of the construct being measured. In some nontechnical usages, a synonym for bias. Refer to Bias, Item.

**Distracter** – A wrong answer in a selected-response item. Refer to Item, Selected-Response Item.

**Domain** – A defined universe of knowledge, skills, abilities, attitudes, interests, or other characteristics.

**Equating Set** – A group of items used to adjust scores on two or more alternate forms of a test so that the scores may be used interchangeably. Refer to Alternate Form, Test.

***ETS Standards for Quality and Fairness*** – A publicly available document offering policy-level guidance to ETS staff on the design, development, and delivery of technically sound, fair, accessible, and useful products and services. Programs are audited for compliance with the *Standards*. Available at no cost from https://www.ets.org/s/about/pdf/standards.pdf. Compare *Standards for Educational and Psychological Testing*.

**Fairness** – For tests, there are many, often conflicting, definitions of fairness. Some definitions focus on equal outcomes for people with the same scores, regardless of group membership. Other definitions focus on proportionate representation of various groups, even if the groups have different average scores. One useful definition of fairness is that a test is fair if any group differences in performance are valid. The existence of group differences in performance does not, by itself, make a test unfair, because the groups may actually differ on the construct being measured. Refer to Bias, Construct, Test, Validity.

**Form** – An edition of a test. Refer to Alternate Form.

**Formative Evaluation** – Tests given during instruction to help shape ensuing instruction. Compare Summative Evaluation.

**Impact** – A difference between groups in percent correct on an item, or in scores, or in passing rates on a test **NOT** adjusted for differences in ability between the groups. Compare Differential Item Functioning (DIF).

**Intended Population** – The test takers for whom a test has been designed to be most appropriate.

**Item** – A test question, prompt, problem, or task. Refer to Task.

**Item Response** – (1) A person's answer to a question. (2) The answer to a question coded into categories such as right, wrong, or omit.

**Item Type** – In some usages, item type is determined by the response mechanism of an item. More generally, the observable format of a test question, problem, or task. For example, a constructed-response item and a multiple-choice item are different item types. Refer to Constructed-Response Item, Multiple-Choice Item.

**Joint Standards** – Refer to *Standards for Educational and Psychological Testing*.

**Key** – The correct answer to a test question, or a listing of the correct responses to a set of test questions. Refer to Test.

**KSA** – Initialism for “knowledge, skill, or ability.” KSA is also used more inclusively to mean “knowledge, skill, or other attribute.”

**Licensing** – The granting by a government agency of permission to practice an occupation or profession, based on evidence that the applicant has the knowledge and skills needed to practice that occupation without endangering the public.

**Linguistic Demands** – The reading or listening ability necessary to comprehend the questions or tasks on a test and the writing or speaking ability necessary to respond.

**Mean** – An arithmetic average. The sum of a set of numbers divided by the number of numbers in the set.

**Modification** – A change to a test or its administration to make the test accessible for a person with a disability. Refer to Accommodation, Construct, Disability.

**Multiple-Choice Item** – A type of selected-response item in which the test taker selects the correct response from a limited number of answer choices (generally four or five). Refer to Selected-Response Item. Compare Constructed-Response Item.

**Noncognitive Test** – A test of attitudes, feelings, beliefs, opinions, personality traits, and the like. Refer to Test. Compare Cognitive Test.

**Norm Group** – A group of test takers used as a basis for comparison. The scores of individual test takers are given meaning by comparison to the distribution of scores in the norm group. Refer to Score.

**Operational Administration** – The use of a test to obtain scores that will be used for their intended purposes. Refer to Score, Test. Compare Pretest.

**Option** – Any of the possible responses to a selected-response item, including both the key and all distracters.

**Overprediction** – When using test scores in a regression equation to predict a criterion such as college grades, overprediction occurs when the predicted criterion measures are higher, on average, for members of a group than the values actually obtained. Refer to Criterion, Regression Equation, Underprediction.

**Performance Item** – An item in which test takers actually do a realistic task rather than answer a question, such as teach a class, parallel park a car, play a particular piece of music, complete a chemistry experiment, repair a clogged fuel injector, perform an appendectomy, land an airplane, or use some software package. Refer to Item, Task.

**Pool** – The set of items from which a test or group of tests will be assembled. Refer to Item, Test.

**Population** – All the members of some defined group, such as third-grade students in the United States. Most populations are too large for every member to be tested. Compare Sample.

**Population Group** – A part of a larger population that is defined on the basis of a characteristic such as gender, race or ethnic origin, training or formal preparation, geographic location, income level, disability, or age.

**Presentation Mode** – Refer to Administration Mode.

**Pretest** – A nonoperational trial administration of items or a test to gather data on item or test characteristics, such as difficulty. Compare Operational Administration.

**Program** – An integrated group of ETS products or services serving similar purposes and/or similar populations. A program is characterized by its continuing character and by the inclusiveness of the services provided.

**Psychometrician** – A specialist in the science of psychological and educational measurement.

**Registration** – The process of enrolling to take a test.

**Regression Equation** – A formula, often of the form $Y = aX + b$, used to estimate the value of a criterion ($Y$), given the value of one or more observed variables ($X$) used as predictors. For example, a regression equation is used to estimate college grade-point average, given high school grade-point average and SAT scores. Refer to Criterion, Overprediction, Underprediction.

**Response Mode** – The procedure used by a test taker to indicate an answer to a question, such as a marking an answer sheet, handwriting, using a mouse, or keyboarding on a digital device. Compare Administration Mode.

**Sample** – A subset of a larger population. For example, a few hundred high schools may be selected to represent the more than 20,000 high schools in the United States. Samples differ in how well they represent the larger population. Generally, the care with which a sample is chosen has a greater effect on its ability to represent a population than does the size of the sample.

**Score** – A quantitative (such as a scale score of 200) or categorical (such as “pass” or “fail”) value assigned to a test taker as the result of some measurement procedure.

**Score Recipient** – A person or institution receiving the scores of individual test takers or summary data for groups of test takers.

**Score Scale** – The set of numbers within which scores are reported for a particular test or program, often, but not necessarily, having a specified mean and standard deviation for some defined reference group. Refer to Mean, Standard Deviation.

**Scoring Rubric** – A set of rules and guidelines for assigning scores to constructed-response or performance items. Generally, there is a description of the attributes of responses associated with each score level. Often rubrics are accompanied by examples of responses at each of the various score levels. Refer to Constructed-Response Item, Performance Item.

**Selected-Response Item** – Item in which test takers select the right answer or answers from a set of choices included in the item. A multiple-choice item is a common type of selected-response item. Refer to Item, Key, Multiple-Choice Item. Compare Constructed-Response Item.

**Specifications** – Detailed documentation of the intended characteristics of a test, including but not limited to the content and skills to be measured, the numbers and types of items, the level of difficulty and discrimination, the timing, and the layout. Refer to Discrimination, Item, Test.

**Standard** – A ruling guide or principle.

**Standard Deviation** – (1) A measure of the dispersion or spread among the numbers in a set of measurements. If the numbers are packed closely together, the standard deviation is small. The further apart from each other the numbers are, the larger the standard deviation will be. (2) Technically, the standard deviation is the square root of the variance. Refer to Variance.

**Standardized Test** – The administration of a test in the same manner to all test takers to allow fair comparison of their scores. Factors such as timing, directions, and use of aids (e.g., calculators) are controlled to be constant for all test takers, including those with disabilities who may require accommodations or modifications. Refer to Accommodations, Disability, Modifications, Score, Test.

***Standards for Educational and Psychological Testing*** – A document published by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). It lists what the publishers consider to be the appropriate ways to develop, use, and evaluate tests. Compare *ETS Standards for Quality and Fairness*.

**Stimulus** – The materials on which an item is based, such as a reading passage, a graph, a political cartoon, a historical document, a video, and so forth. Refer to Item.

**Studied Group** – A population group sampled in evaluations of the performance of a test for people in different groups, particularly with respect to fairness. Refer to Population Group.

**Summative Evaluation** – Test of a student at the completion of a unit of instruction. Compare Formative Evaluation.

**Task** – The behaviors that test takers are required to exhibit to provide evidence of some particular knowledge, skills, or other attributes (KSAs); the materials presented to test takers; and the manner in which those materials are presented. A task can include either a single item or several related items, as well as all of the relevant features of those items and of any directions and stimuli that are used. Refer to Item.

**Technology-Enhanced Items (TEIs)** – TEIs are items for which computers or other digital devices are used to allow various item presentation and response modes that are difficult or impossible to present on paper-based tests with answer sheets. TEIs include both selected-response items and constructed-response items. Technology-enhanced items can use realistic simulations, interactive displays, or video and audio clips as stimuli. Test takers can be asked to use search engines and word processors. The response modes are highly varied and can include constructing graphs or histograms, selecting cells in a matrix, selecting answers from lists of choices, moving one or more objects to a given set of locations (drag and drop), entering numbers or equations, entering brief or extended text, providing an oral response, marking locations on a number line, selecting parts of images or blocks of text, or running a simulation. Refer to Item, Constructed-Response Item, Selected-Response Item.

**Test** – A systematic sample of behavior taken to allow inferences about an individual's knowledge, skill, ability, or other attribute.

**Underprediction** – When using test scores in a regression equation to predict a criterion such as college grades, underprediction occurs when the predicted criterion measures are lower, on average, for members of a group than the values actually obtained. Refer to Criterion, Overprediction, Regression Equation.

**User** – Individual or institution making decisions on the basis of test scores.

**Validity** – The extent to which inferences and actions made on the basis of test scores are appropriate and justified by all of the theoretical and empirical evidence bearing on what a test is actually measuring. It is the most important aspect of the quality of a test. Validity refers to how the scores are used rather than to the test itself.

**Variable** – An attribute that can take on different values, such as test scores, grade-point averages, family incomes, ages, and weights.

**Variance** – (1) Generally, a label for differences among scores. "Sources of variance," for example, refers to causes of the differences among test takers' scores. (2) Technically, a statistic characterizing the magnitude of the differences among a set of measurements. Specifically, it is the average squared difference between each measurement and the mean of the measurements.

# 16.0 Appendix 1: Plain Language

The purpose of this appendix is to help you make the language in ETS tests and associated materials as clear and as comprehensible as is consistent with validity. In some cases, clients have specific style guides or other policies related to language use. The *ETS Guidelines for Using Plain Language* is not intended to conflict with client preferences.

While some test takers may especially benefit from the use of plain language (e.g., those with limited knowledge of English, those with disabilities related to language processing, those who are not strong readers), use of plain language is not a testing accommodation. It is a practice that seeks to improve validity and fairness for all test takers by reducing construct-irrelevant score differences.

## 16.1 Application

These guidelines apply to all test takers, to all construct-irrelevant elements of tests, and to all associated test materials (registration bulletins, score reports, etc.).

Language that is part of the construct being tested should be as complex and as challenging as needed for valid measurement. The need to assess the construct thoroughly and accurately must always be placed above the desire to make language simpler.

The following are some specific examples of instances where prioritizing plain language is **<u>NOT</u>** appropriate:

- In reading-comprehension tests, the stimuli should be governed by the construct being tested. A reading test for college admission should contain entry-level college text.
- Subject-matter tests use specialized vocabulary and language structures that are part of the subject matter. It is entirely appropriate, therefore, to use vocabulary unfamiliar to the general public, such as “ontogeny” on a biology test or “metonymy” on a literature test.
- Historical documents may use archaic and difficult language if the ability to understand such documents is part of the intended construct.
- In assessments of language proficiency (e.g., Test of English as a Foreign Language™ [TOEFL®], AP® Spanish), the level of complexity and challenge of the stimuli and test items should be entirely determined by the construct being assessed.

In all cases, however, the construct-irrelevant aspects of the test material should use the simplest language that is consistent with validity. In many cases, this means that the stimulus should be as complex and challenging as necessary to assess the construct, while items should be as clear and comprehensible as possible.

## 16.2 Guidelines for Plain Language

Writing in a clear and accessible way requires care and a clear understanding of what you are trying to say. In striving to write in the most inclusive language possible, consider the audience and the construct being assessed and continually look for ways to increase clarity and improve comprehension. The ideas presented in this section are guidelines to help meet that goal rather than as rules to be followed strictly. (For more detailed guidance regarding plain language, see https://www.plainlanguage.gov/.)

## 16.3 Paragraphs

Try to use short, clear paragraphs with one main idea.

- In expository writing, state the main idea of a paragraph in the first or second sentence.
- Make the connections between paragraphs clear. Do the paragraphs represent steps in a sequence? Does a following paragraph offer examples of a concept described in the preceding paragraph? Does it offer a contrasting point of view?

## 16.4 Sentences

When appropriate, use short, simple sentences with a subject-verb-object structure. Bear in mind, however, that sentences that are too short and choppy can sometimes impede communication. Be guided by the ideas that need to be expressed.

- Take care in using relative clauses (e.g., the underlined clause “that I am reading” in the sentence “The book <u>that I am reading</u> is interesting”). While relative clauses can be an effective means of representing complex ideas in a single sentence, their overuse can make sentences difficult to follow.
- Make the referent of a pronoun as clear as possible. Usually, the referenced noun should be the closest one (of the same grammatical number) before the pronoun. If there is any possibility of ambiguity, repeat the noun rather than use a pronoun (e.g., for “When Connie joined Marta in the project, she did not know that she would be so competitive,” instead write “When Connie joined Marta in the project, she did not know that Marta would be so competitive.”)
- Use transition words (e.g., “however,” “first,” “next”) whenever they increase clarity. It is acceptable in directions to start sentences with conjunctions such as “and,” “but,” or “however.”

### 16.5 Vocabulary

Use vocabulary that is widely understood by test takers. Whenever possible, use common words rather than less common synonyms (e.g., “walk” rather than “ambulate”).[19]

- Be concise, but do not reduce clarity to save a few words.
- Try to use specific, concrete words rather than more abstract words.
- Avoid the use of foreign expressions that may be less familiar than common English equivalents (e.g., “C’est la vie” versus “That’s life”) unless such foreign expressions are construct relevant or in historical texts.
- Avoid colloquial and idiomatic expressions, including slang or dialect. Such language can be understood differently by test takers from different backgrounds and is likely to be particularly challenging for people who are English-language learners.
- Avoid construct-irrelevant figurative language. Be straightforward and direct.
- Be consistent in the use of terminology. Avoid using different words to refer to the same thing (e.g., “subject,” “discipline,” “field”).
- Avoid acronyms, initialisms, and abbreviations, unless they are more familiar than the full terms (for example, “DNA” is likely to be more accessible than “deoxyribonucleic acid”). When using acronyms, initialisms, or abbreviations that might be unfamiliar to some test takers, explain them or give the full term on first use.
- Avoid long noun phrases. Noun phrases with multiple modifiers (e.g., “ethernet hub cable connection”) are often hard to process because it can be unclear whether a given word is being used as a noun or a modifier.
- Avoid using words with multiple meanings in contexts where the meaning might not be clear (unless assessing words with multiple meanings is important to the construct).
- When using words in a part of speech that is not common for the word (e.g., “foot” as a verb), take care to ensure that the context makes the intended meaning clear.
- Use personal pronouns when they help with communication. When appropriate, in directions address the reader as “you” rather than using a more abstract, impersonal reference such as “one.”

[19] Particularly at the K–12 level, vocabulary guidelines such as Mogilner & Mogilner (2006) *Children’s Writers’ Word Book* and Taylor et al., *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies* (1989) can be useful resources.

- If a passage contains challenging vocabulary that is not part of the construct and that cannot be edited, consider adding the words to a glossary or footnoting the difficult words. Footnotes should be used, however, only when they are customary to the program and when the test takers can be expected to be familiar with footnotes.

### 16.6 Verb Forms

- Use the simplest verb forms that will clearly communicate your meaning. Try to use the simple past, the simple future, and the simple present whenever possible and to use more complex verb forms only when necessary.
- Use active voice rather than passive voice unless there is a clear advantage to using the passive. For example, use "Toni Morrison wrote *The Bluest Eye*" instead of "*The Bluest Eye* was written by Toni Morrison."
- Use the imperative mood to give directions. For example, "Select the best answer to each question" is clearer than "Each student should select the best answer to each question."

### 16.7 Layout and Formatting

Use layout and formatting to make the organization of your writing clear to the reader and easy to understand. Well-designed headings and graphic arrangement can help the reader to recognize the relative importance of information and the order in which it should be considered.

- Use numbered or bulleted lists for directions and other material that can be better comprehended in list form.

### 16.8 Some Particular Issues for Test Items

The previously discussed guidelines apply to all types of writing. What follows are guidelines that are specific to test items. Some of them are more relevant to the test design stage (at which practices such as standard wording of stems are likely to be established), while others apply to ongoing item writing and review.

### 16.9 Stems

The stem is the part of the test item that poses a question or otherwise sets a task for the test taker. Stems should present the task as clearly and precisely as is consistent with valid measurement.

- Consider the strengths and weaknesses of both closed stems and open stems in multiple-choice items. (Closed stems ask a complete question. Open stems state an incomplete sentence to be completed by one of the answer choices.) Closed stems are often preferred because by presenting a complete question they may make the student's task clearer. However, open stems sometimes allow a more concise presentation of the task.

- If multiple-sentence stems are an acceptable style in the program, consider breaking up long stems into separate sentences. For example, "If $S$ represents the number of sheep a farmer owned, which of the following number sentences represents the number of sheep the farmer had after selling 3 of the sheep?" This stem can be presented more clearly as a series of simple sentences: "A farmer had $S$ number of sheep. The farmer sold 3 of the sheep. Which number sentence represents how many sheep the farmer has now?"
- Try to minimize the use of negative stems. Where they are used, there should be appropriate emphasis (such as "<u>**NOT**</u>" or "<u>**EXCEPT**</u>" in all capital letters, with the words in boldface and underlined) to reinforce that the stem is negative.

### 16.10 Examples

The following examples have been selected to show how the language of test items can be modified to increase comprehension and ease of reading without affecting the construct being assessed.

***College Placement—Critical Reading***

| **Less Accessible** | **More Accessible** |
|---|---|
| From the passage above, one can infer that the author is using the word "panacea" to mean which of the following? | As used in the passage, the word "panacea" means |

*Comment:* The less accessible version introduces extraneous language. The more accessible version is succinct.

***Elementary Mathematics***

| **Less Accessible** | **More Accessible** |
|---|---|
| If a single card is to be chosen from the group without looking, what is the probability that it will be a blue card? | A student will pick one card from the group without looking at it. What is the probability that the student will pick a blue card? |

*Comment:* Rather than posing a hypothetical question, using a simple and relatable scenario in which a student is the subject and then breaking the stem into two sentences help to present the task more clearly.

| Less Accessible | More Accessible |
|---|---|
| When Ms. Johnson pulled her car into the parking garage, she received a ticket stamped with the time 11:12 A.M. When she left the garage that afternoon, the time was 2:15 P.M. What was the total length of time that Ms. Johnson's car was in the parking garage? | Anika went into the library at 11:12 A.M. She left the library at 2:15 P.M. the same day. How long was she in the library? |

*Comment:* The less accessible version uses a context likely to be unfamiliar to many elementary students and to many rural students. That version is also unnecessarily wordy. The more accessible version uses a simpler context to measure the same construct.

***Social Studies — High School***

| Less Accessible | More Accessible |
|---|---|
| The development of the concept of interchangeability of parts and the introduction of the assembly line in industrial manufacturing allowed the owners of factories to make more efficient use of . . . | The assembly line and the interchangeability of parts allowed factory owners to make more efficient use of . . . |

*Comment:* The less accessible version begins with a long introductory noun phrase that contains several abstract words (e.g., “development,” “concept,” “introduction”). The more accessible version is a simpler means of assessing the same construct.

# 17.0 Appendix 2: Abridged List of Guidelines for Fairness

## 17.1 Purpose

This abridged list of guidelines is **NOT** a stand-alone document to be used as a substitute for the *GDFTC*.

The purpose of this highly abridged listing of fairness guidelines is to help ETS staff use the guidelines in their daily work. The body of the *GDFTC* includes explanations, discussions, caveats, examples, and additional material to help users understand and interpret the guidelines appropriately. Those features are intended to help users learn about the guidelines but they may interfere with use of the document as a convenient reference while working on test materials. Therefore, use the following abridged list of guidelines as a work aid only after becoming familiar with the information in the body of the document.

## 17.2 Meanings of Fairness for Tests

Fairness in the context of assessment can usefully be defined as the extent to which inferences and actions based on test scores are valid for diverse groups of test takers.

### 17.3 Groups to Consider

Pay special attention to groups that are discriminated against. It is also necessary to consider intersectionality (the combination of two or more groups) when evaluating fairness.

### 17.4 Interpreting the Guidelines

Consider the following factors when deciding whether material complies with the guidelines.

***Importance for Validity.*** Any material that is important for valid measurement—and for which a similarly important but more appropriate substitute is not available—may be acceptable for inclusion in a test, even if it would otherwise be out of compliance with the guidelines.

### 17.5 Interpretation of Guidelines

| More Lenient Interpretation | More Strict Interpretation |
| --- | --- |
| • No individual scores<br>• No important consequences<br>• Measurement of specific subject-matter content<br>• Older test takers<br>• More experienced test takers<br>• Externally owned copyrighted material<br>• Brief mention<br>• Indirect allusion<br>• Lenient interpretation desired by client<br>• Instructional material<br>• Teacher guidance<br>• Need for authentic stimulus | • Individual scores<br>• Important consequences<br>• Measurement of skills that could be based on many different contents<br>• Younger test takers<br>• Less experienced test takers<br>• Material written by ETS<br>• Extended discussion<br>• Direct mention<br>• Strict interpretation desired by client |

### 17.6 General Principles for Fairness

- Measure the important aspects of the intended construct.
- Avoid construct-irrelevant barriers to the success of test takers.
- Provide scores that support valid inferences about diverse groups of test takers.
- Treat all test takers respectfully and impartially.

### 17.7 Construct-Irrelevant KSA Barriers to Success

Construct-irrelevant knowledge, skill, or ability barriers to success may arise when KSAs that are not part of the construct are required to answer an item correctly.

***Contexts***
Contexts should engage test takers rather than puzzle or distract them. The information required to answer the items correctly should either be common knowledge among the intended test takers or be available in the stimulus.

***Disabilities***
Do not use test items (including stimuli) for which a correct response requires personal experiences that may not be available to test takers with disabilities. Consult with ETS's accessibility experts in the event that such items are critical for valid measurement.

***Language***
Use the simplest and clearest language that is consistent with valid measurement. Avoid requiring construct-irrelevant knowledge of specialized vocabulary or figurative language.

***Regionalisms***
Do not require knowledge of words, phrases, or concepts that are more likely to be known by people in some regions of the United States than by people in other regions, unless this knowledge is important for valid measurement.

***Religion***
Do not require construct-irrelevant knowledge about any religion to answer an item. If the knowledge is part of the construct, use only the information about religion that is important for valid measurement.

***Specialized Knowledge***
Avoid requiring construct-irrelevant, specialized knowledge to answer an item correctly.

***Translation***
Translation alone may be insufficient for many test items. The content of items must be adapted for the culture of the country in which the items will be used.

***Unfamiliar Item Types***
Make clear to the test taker what action is needed to respond to the item. Using the computer should not be a construct-irrelevant source of difficulty.

Be consistent in the way that items of the same or very similar item type are presented. Items that may appear to be similar but are actually different should have directions that call attention to their differences.

***United States Culture***
Do not require a test taker to have specific knowledge of the United States to answer an item, unless the item is supposed to measure such knowledge.

Do not assume that all test takers are from the United States.

In tests that will be used worldwide, avoid construct-irrelevant images of people posed, dressed, or behaving immodestly.

## 17.8 Construct-Irrelevant Emotional Barriers to Success

Construct-irrelevant emotional barriers to success arise when language or images cause strong emotions that may interfere with the ability of some groups of test takers to respond to an item. Passages about some group of people need careful scrutiny for any construct-irrelevant material that might plausibly cause a negative reaction.

***Topics to Avoid***

Some topics (e.g., abortion, atrocities, rape, torture) are so controversial, so demeaning, so inflammatory, or so upsetting that they are best avoided unless they are important for validity.

***Topics Requiring Care***

- **Advocacy.** Do not use test content to advocate for any contested cause or ideology or to take sides on any controversial issue unless doing so is important for valid measurement.
- **Avatars.** Show diverse genders, races, and ethnicities, or use unrealistic avatars that do not show those characteristics.
- **Biographical Material.** Avoid construct-irrelevant focus on individuals who are associated with offensive topics or controversial activities. Avoid construct-irrelevant focus on live celebrities.
- **Brand Names.** Avoid construct-irrelevant brand names. Communications other than test materials may mention brands as appropriate.
- **Conflicts.** Unless important for validity, do not take the point of view of one of the sides in a conflict in which test takers may sympathize with different factions.
- **Cryptic References.** Be alert for cryptic references to anti-Semitism, drugs, gangs, homophobia, sex, White supremacy, racism, and other unsuitable topics. Check the meanings of unfamiliar names, numbers, images, or words that appear to be arbitrary, out of place, or strange.
- **Disability.** Avoid negative or derogatory references to people with disabilities. Avoid the implication that people with disabilities are less valued members of society than are members of the general population.
- **Evolution.** Avoid items or stimuli concerning the evolution of human beings and the similarities of human beings to other primates unless important for valid measurement.
- **Group Differences.** Avoid unsupported generalizations about group differences. Do not state or imply that any groups are superior or inferior to other groups with respect to

subjectively evaluated traits. Do not overrepresent members of a group as showing irrational or criminal behavior.

- **Humor, Irony, and Satire.** Avoid construct-irrelevant humor, irony, and satire.
- **Luxuries.** Avoid depicting construct-irrelevant situations that are associated with excessive spending on what many members of the test-taking population would consider luxuries.
- **Maps.** Unless important for valid measurement, avoid showing maps of disputed areas indicating that the area belongs to one of the parties in the dispute.
- **Mistreatment of Groups.** Unless important for validity, avoid materials that depict any group that is or was passively suffering the effects of prejudice, being harmed or exploited by a supposedly superior group, benefiting from contact with a supposedly superior group, or emulating a supposedly superior culture.
- **Personal Questions.** Unless important for validity, avoid asking test takers to respond to excessively personal questions regarding themselves, their family members, or their friends.
- **Religion.** Avoid construct-irrelevant material that focuses on any religion, any religious group, any religious holidays, any religious practices, any religious beliefs, any conflicts between religions, or anything closely associated with religion (including the creation stories of various cultures). Also avoid material on the lack of religion, agnosticism, or atheism.
- **Role Playing.** Do not ask test takers to take on construct-irrelevant roles that might cause test takers emotional distress or be counter to their strongly held beliefs.
- **Sexual Behavior.** Avoid construct-irrelevant double entendres, sexual innuendo, and explicit descriptions of human sexual acts.
- **Slavery.** Avoid construct-irrelevant materials about slavery. A brief mention of slavery may be acceptable if it is clear that the passage is about something else. Though “slave” is still an acceptable term, “enslaved person” is preferred, though note that “enslavement” is not an acceptable term for the general term “slavery.” “Slaveholder” is preferred to “slave owner.” Authentic materials that use the older terms are acceptable.
- **Stereotypes.** Avoid stereotypes (both negative and positive) in language and images unless important for valid measurement. If some group members are shown in traditional roles, other members of the group should be shown in nontraditional roles.
- **Unstated Assumptions.** Avoid material based on underlying assumptions that are false or that would be inappropriate if the assumptions had been stated.

Be careful using the word “we” unless the people included in the term are specified.

- ***Violence and Suffering.*** Do not focus on violent actions, on violent crimes, on the detailed effects of violence, or on suffering unless important for valid measurement. Violence and suffering are too common to exclude them completely from all material.

- ***Visual Material.*** Do not use visual material without a clear purpose for doing so. Use the simplest images that are consistent with valid measurement and the need for authenticity. Avoid unnecessary visual clutter whenever possible. Consider the difficulty of describing the image in words when selecting visual materials. Scrutinize the background of visual material as well as the foreground. Magnify the image as necessary. Unless important for valid measurement, avoid visual material that depicts content out of compliance with the guidelines in this document. Translate language as necessary for evaluation.

### 17.9 Construct-Irrelevant Physical Barriers

#### *Requirements*

Tests and related materials must be digitally accessible and conform to standards. Paper-delivered tests must be amenable to adaptation into alternate formats.

#### *Types of Physical Barriers*

Construct-irrelevant physical barriers to success occur when aspects of tests **<u>not</u>** important for validity interfere with the test takers’ ability to attend to, see, hear, or otherwise perceive the items or stimuli and/or to enter a response to the item.

- **Essential Aspects.** Some aspects of tests are important or essential for validity and no acceptable substitute exists. They are, therefore, acceptable; consult with ETS accessibility experts for proactive remediation and solutioning prior to the finalization of test design in the event that essential aspects of a test might cause difficulties for test takers with disabilities.

- **Helpful Aspects.** Some physical aspects of various tests are helpful for measuring the intended construct. Helpful aspects of a test may be retained if they are accessible and allow people with disabilities to interact with and respond appropriately to the item types.

- **Unnecessary Aspects.** Avoid material that is needlessly complicated, hard to discern, or confusing.

### 17.10 Appropriate Terminology for Groups

In authentic historical and literary material, some violations of the guidelines may be inevitable.

| Group | Preferred | Conditional | Avoid |
|---|---|---|---|
| All | • Names that group members prefer<br>• Names as adjectives (e.g., Black people) | Names as nouns (e.g., "the Blacks"); OK to use sparingly after first use as an adjective | Derogatory names, even if used by some members of that group |
| People who are African American | African American, Black | Colored, Negro, Afro-American (OK only in historical material or names of organizations) | The word "black" as a negative adjective (e.g., "black day")<br><br>Note: The term "people of color" and the acronym BIPOC are not synonyms for "Black" or for any other specific group but, rather, refer to a general mixed group of non-White ethnic groups. |
| People who are Asian American | • Asian American, Pacific Island American, and Asian/Pacific Island American<br>• Specific country names are preferred (e.g., "Chinese American," "Japanese"). | N/A | The term "Oriental" (except in literary or historical material or as part of names of organizations) |
| People with disabilities | • Person-first terminology (e.g., "person who is blind")<br>• Objective, neutral phrasing (e.g. "uses a wheelchair") | Disability-first construction may be used after person-first terminology. | • Negative terms (e.g., "confined to a wheelchair")<br>• Patronizing terms (e.g., "special," "physically challenged")<br>• The term "handicap" to refer to a disability |
| People who are blind | The terms "blind," "visually impaired" (for varying degrees of vision loss) as adjectives | "Blind" as a noun OK only in names of organizations or institutions | N/A |

| Group | Preferred | Conditional | Avoid |
| --- | --- | --- | --- |
| People with a learning or cognitive disability | • Cognitively disabled, cognitively impaired, developmentally delayed, developmentally disabled<br>• Down syndrome | Disability-first terminology is generally preferred for autism (e.g., autistic person) | Down's syndrome, Mongoloid, retarded, slow |
| People who are deaf | • The terms "deaf" and "deaf and hard of hearing" as adjectives<br>• Use uppercase for "Deaf" referring to the culture lowercase when referring to the disability | "Deaf" as a noun OK only in names of organizations or institutions | Deaf and dumb, deaf mute, hearing impaired |
| People with a motor disability | Motor disability, motor impairment | "Paraplegic" and "quadriplegic" are OK as adjectives but not as nouns. | The term "spastic" is unacceptable when used to describe a person. |

| Group | Preferred | Conditional | Avoid |
|---|---|---|---|
| People of different genders, sexes, and sexual orientations | • Refer to people of all genders, sexes, and sexual orientations in parallel terms.<br>• Use the term a person prefers, if known.<br>• "Gay" as an adjective,<br>• "LGBTQ+" seems most appropriate. Be sure to define the initialism the first time you use it, and be sure it is representative of the groups about which you are writing.<br>• Use the pronoun preferred by the person,<br>• Use plural constructions, and constructions that avoid pronouns.<br>• "Ms." for women | • Gender or sex as a distinguishing feature<br>• "Queer" if OK with program<br>• Singular "they," if OK with program<br>• "Male" and "female" as adjectives<br>• Use "ladies" only when men are called "gentlemen."<br>• Use "wives" only when men are called "husbands."<br>• "Mrs." if the person prefers or if it is used with "Mr." (e.g., "Mr. and Mrs. Ruiz")<br>• "Mx." if OK with program<br>• Homosexual only in a scientific, literary, or historical context | • "Sexual preference"<br>• "Normal" and "abnormal"<br>• "The opposite sex"<br>• Assumption that gender or sex is binary or that marriage is only between a man and a woman<br>• "Gay" as a noun<br>• Construct-irrelevant references to appearance<br>• "Male" and "female" as nouns<br>• "He" or "man" to refer to all people<br>• "He or she" and "his or hers"<br>• Occupations ending in "man"<br>• Jobs by gender (e.g., male nurse, poetess)<br>• Gender-specific words for objects |
| People who are Hispanic American | • "Latino American" (for men and mixed-gender groups), "Latina American" (for women), "Hispanic American"<br>• Specific group names such as "Cuban American" or "Cuban" are preferred<br>• Chicano, Chicana, and Latinx are acceptable if OK with program. | • Specific group names such as "Cuban American" or "Cuban" are preferred<br>• Chicano, Chicana, and Latinx are acceptable if OK with program. | N/A |

| Group | Preferred | Conditional | Avoid |
|---|---|---|---|
| People who migrate to the United States | • The phrase "people who migrate to the United States"<br>• Immigrant, migrant, undocumented immigrant | N/A | • The phrase "illegal alien"<br>• The word "illegal" as a noun |
| People who are members of more than one racial/ethnic group | Biracial, multiracial, people of color | Minority, majority | The phrase "colored people" (except in historical or literary material or in the name of an organization) |
| People who are Native American | • Native American, American Indian, Native American,<br>• Indigenous people<br>• Members of the First Nations (Canada)<br>• Specific group names that people call themselves | "Tribe" is OK if used by members of the group. Some prefer "nation." | • Eskimo<br>• Brave, buck, squaw |

| Group | Preferred | Conditional | Avoid |
| --- | --- | --- | --- |
| People who are nonnative speakers of English | • "English-language learner (ELL)," "English learner (EL)" used for K–12 students<br>• "English as a second language (ESL)" refers to people who are learning English in an English-speaking environment<br>• "English as a foreign language (EFL)" refers to people who are learning English in a non-English-speaking environment<br>• Use "ESL," "EFL," and "LEP" as adjectives, not as nouns.<br>• Use "ELL" and "EL" as adjectives.<br>• For any of these abbreviations, the first appearance in text, whether as a noun or an adjective, should be accompanied by the spelled-out term in parenthesis. All instances thereafter can be abbreviated, even when the abbreviation is used to refer to people. | • "Limited English proficient (LEP)" is generally limited to legislation<br>• "ELL" and "EL" are OK as nouns after they have been used as adjectives. | "ESL," "EFL," and "LEP" as nouns |

| Group | Preferred | Conditional | Avoid |
|---|---|---|---|
| People who are older | • Specific ages or age range<br>• Older people<br>• Person with dementia | N/A | • “Aged” and “Elderly” as nouns<br>• Euphemisms such as “senior citizens” or “seniors”<br>• Derogatory age-related adjectives such as “senile” |
| People who are White | • White<br>• European American | • Caucasian<br>• Anglo American | The word “white” as a positive adjective |

### 17.11 Representation of Diversity

Represent diversity in tests that mention people or show people in images and among authors of materials.

***People with Disabilities***
If suitable for the subject matter, try to have about 10 to 15 percent of the items that depict people include people with disabilities.

***Gender Balance***
In tests that predominantly measure skills, genders should be represented in comparable ways and numbers. The gender balance of tests that predominantly measure content or an occupation should be appropriate to the subject matter or occupation. Strive for some diversity when possible.

***People Who Are LGBTQ+***
People may be identified by sexual orientation in tests when it is construct relevant to do so. Identify people by sexual orientation in tests for purposes of representing diversity with the approval of the client.

***Racial and Ethnic Balance***
In tests that predominantly measure skills, try to have about at least one-third of the items that mention or show people represent people from what are commonly considered to be minority groups in the United States or people from the countries of origin of those groups.

In tests of specific subject matter or occupations, try to meet the representational goals given above to the extent suitable for the subject matter or occupation. Strive for some diversity when possible.

***Societal Roles***
If it is possible to do so in the materials for a test, demonstrate that people in different groups are found in a wide range of societal roles and contexts.

# 18.0 Additional Guidelines for Fairness of NAEP and K–12 Tests

These guidelines for NAEP and K–12 tests are in addition to, not a replacement for, the guidelines that apply to all ETS tests.

**Requirements for NAEP.**

According to the National Assessment Governing Board, NAEP items must be secular, be neutral and non-ideological, and avoid sensitive topics.

Older items used to measure trends over time are generally acceptable even if they do not comply with all current fairness guidelines.

**K—12 Assessments.**

K–12 assessments include tests commissioned or selected by consortiums of states, individual states, cities, or school districts for use in their classrooms from kindergarten through the end of high school.

Different K–12 clients may have different fairness requirements, and fairness requirements change over time. Check with the responsible assessment director for the current fairness requirements of the client.

***Emotionally Charged Topics.***
Avoid construct-irrelevant discussions of topics that may be excessively emotionally charged for K–12 students, such as unusual physical attributes, dissension, disasters, family problems, and violence.

***Individual and Group Names.***
Ask the program assessment director to identify the program's preferred terms for groups and which names or other terms or phrases that the program prefers to refer to individuals.

***Offensive Topics.***
Avoid construct-irrelevant topics that may be offensive to particular groups in a jurisdiction, such as smoking, gambling, and the occult.

***Controversial Topics.***
Do not promote or defend personal or political values in K–12 test materials. Maintain a neutral stance on controversial issues that are particularly troublesome in some jurisdictions (e.g., deforestation, gun control, vaccination).

***Inappropriate Behavior.***
Do not use material that models or reinforces inappropriate student behaviors (e.g., lying, stealing, entering homes of unknown adults, using weapons, breaking laws)