Measuring the Power of Learning.®

ETS Guidelines for Fair Tests and Communications

By Educational Testing Service

Contents

PREFACE

One of my tasks as Senior Vice President and General Counsel at ETS is to serve as the officer with responsibility for the guidelines that help to ensure the fairness of ETS tests and of ETS communications in all media. The guidelines are an essential tool in accomplishing the ETS mission "to help advance quality and equity in education by providing fair and valid assessments, research and related services."

Reviews for the fairness of materials have been carried out at ETS on a voluntary basis since the late 1960s. The reviews became mandatory in 1980 when the first version of these written guidelines was issued. Since that time, the guidelines have been updated approximately every five years, on average. As societal views of fairness have changed, and as more has been learned about fairness, the guidelines have become increasingly inclusive and comprehensive. Our experience with the use of guidelines for fairness for more than three decades has helped to shape the most recent version.

I am pleased to issue the 2015 edition of the *ETS Guidelines for Fair Tests and Communications*. It is my hope that 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.

Glenn Schroeder

Senior Vice President and General Counsel

Educational Testing Service

ETS Guidelines for Fair Tests and Communications

INTRODUCTION

Purpose. The primary purpose of the *ETS Guidelines for Fair Tests and Communications* (*GFTC*) is to enhance the fairness of items, tests, and communications created by Educational Testing Service (ETS). The *GFTC* is intended to help developers and reviewers of ETS materials to

- obtain a better understanding of fairness,
- take fairness into account as materials are designed,
- avoid the inclusion of unfair content as materials are developed,
- find and eliminate any unfair content as materials are reviewed,
- represent diverse people in materials, and
- educe subjective differences in decisions about fairness.

Intended Users. The *GFTC* is written for the people who design, develop, and review ETS items, tests, and related materials in any language or medium, such as cognitive tests, noncognitive tests, equating sets, pretests, questionnaires, surveys, test bulletins, test descriptions, and test-preparation materials. The *GFTC* is also written for the people who conceptualize, generate, and review ETS communications in any language or medium, such as books, films, journal articles, marketing materials, news releases, posters, presentations, proposals, research reports, videos, and Web pages.

Use of the *GFTC* is not limited to ETS staff and consultants, however. The *GFTC* is copyrighted, but it is not confidential. ETS encourages use of the concepts discussed in the *GFTC* by all who wish to enhance the fairness of their tests and communications. (The document may be downloaded at no cost from the Fairness page on ETS's website, www.ets.org/about/fairness.

Overview. Following this introduction, the *GFTC* includes the following topics:

- Fairness for Tests and Communications,
- Reasons for Using the Guidelines,[1]

[1]Guidelines and GFTC (in italics) refer to the document ETS Guidelines for Fair Tests and Communications.

- Applications of the Guidelines,
- Interpretations of the Guidelines,
- General Principles for Fairness,
- Guidelines Regarding Construct-Irrelevant Cognitive Barriers,
- Guidelines Regarding Construct-Irrelevant Affective Barriers,
- Guidelines Regarding Construct-Irrelevant Physical Barriers,
- Guidelines Regarding Appropriate Terminology for Groups,
- Guidelines Regarding Representation of Diversity,
- Additional Guidelines for the Fairness of NAEP and K–12 Tests, and
- Additional Fairness Actions.

The appendices include:

1. Guidelines for Using Accessible Language, and
2. Abridged List of Guidelines for Fairness.

FAIRNESS FOR TESTS AND COMMUNICATIONS

Many Definitions. Designing and developing fair tests and communications requires an understanding of what is meant by "fair" in those contexts. Discussions of fairness in communications have generally focused on the avoidance of sexist language, offensive content, and stereotypes; the use of correct terminology for groups of people; and the representation of diversity. Authors, editors, and reviewers who are familiar with such commonly referenced sources for writers as the *Associated Press Stylebook* (Associated Press [AP], 2015), the *Chicago Manual of Style* (University of Chicago Press, 2010), or the *Publication Manual of the American Psychological Association* (American Psychological Association [APA], 2010) will find the *Guidelines* to be consistent with the sections of those documents that deal with the fairness of communications.

The word "guidelines" in Roman font refers to the individual instructions within the document.

Because a test is a form of communication, there is general agreement that tests should at least follow the guidelines for fairness appropriate for all communications. There is also general agreement that the fairness of tests is more complex than the fairness of other communications. There is not, however, a universally accepted definition of fairness in the context of assessment (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p.49). Psychometricians have proposed a number of different, sometimes contradictory, views of fairness. (For descriptions of various definitions, see e.g., Camilli, 2006; Zieky, 2013.) The definitions are, however, of little direct use in the design and development of tests because the definitions focus on the outcomes of using completed tests for prediction and selection.

Validity. For test designers, developers, and reviewers the most useful definition of fairness in assessment is based on validity. Validity is the most important indicator of test quality. Messick (1989, p.13) 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" (emphasis in the original). More simply, validity is the extent to which a test measures the right knowledge, skills, or other attributes (KSAs) in the right way and provides scores that function appropriately.

The "right KSAs" are more formally referred to as the *construct* that a test is intended to measure. An important aspect of validation is the collection of evidence to show that the intended construct is appropriate. Any KSA that is part of the construct is *construct relevant.* Any KSA that is not part of the construct is *construct irrelevant.* Measuring the KSAs in the "right way" requires measuring a suitable sample of the construct-relevant KSAs while minimizing the inadvertent measurement of any construct-irrelevant sources of score differences. In the case of perfect validity, all of the differences among test takers' scores would come from construct-relevant sources. Perfect validity is impossible, but as the proportion of the score differences caused by construct-relevant sources increases, validity increases.[2]

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

The extent to which communications meet their intended purposes for their intended audiences is analogous to validity in tests. Material identified as inappropriate for tests may also interfere with the ability of a publication to meet its intended purpose for its intended audience. For example, language that is more difficult than necessary to meet the purpose of a test or of a communication will make the test less fair and the communication less effective. Material that may offend test takers may also offend the audiences for other communications.

Fairness. According to the S*tandards for Educational and Psychological Testing* (AERA, APA, NCME, 2014, p. 49), "fairness is a fundamental validity issue." If a construct-irrelevant source of score differences affects all test takers to about the same extent, validity is diminished. If a construct-irrelevant source of score differences affects some group of test takers (e.g., women) more than some other group of test takers (e.g., men), then fairness as well as validity is diminished. Therefore, 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.

Differences in Difficulty. Many people believe that an item that is harder for some group than it is for other groups is not fair for the lower-scoring group. Such items are fair, however, if the differences in difficulty are caused by construct-relevant aspects of the item. Psychometricians disagree about many things concerning fairness, but "if the members of the measurement community agree on any aspect of fairness, it is that score differences alone are not proof of bias" (Cole & Zieky, 2001, p. 375). For example, if an item were intended to measure reading comprehension in Spanish, score differences between native speakers of Spanish and Spanish- language learners would be valid and fair if the two groups really differed in their ability to comprehend Spanish text. The item would be unfair, however, if the score differences between the groups of test takers had been caused by construct-irrelevant aspects of the item.

REASONS FOR USING THE GUIDELINES

Increased Validity and Effectiveness. Following fairness guidelines is not an exercise in political correctness. It is a way to increase validity because fairness is essential for validity. Shepard (1987, p.179) very concisely defined bias as "invalidity." Practices that reduce bias, such as using the *GFTC* to reduce irrelevant and potentially offensive material, will increase validity in the context of tests and effectiveness in the context of communications.

Respect for People. Using the *GFTC* helps to ensure that ETS tests and communications show respect for diverse groups of people, are sensitive to their feelings, avoid content and images that are insulting or demeaning, and are as free as possible of irrelevant obstacles. Using the *GFTC* also helps to ensure that different groups of people are represented and that people are not unnecessarily angered, irritated, or alienated by the contents of ETS tests or communications.

ETS Mission. Use of the *GFTC* helps ETS meet the goals that it has set as part of its mission statement. The ETS mission is**,** in part, "to help advance quality and equity in education by providing fair and valid assessments, research, and related services." Because the *GFTC* focuses on ways to enhance validity and fairness, its use supports the ETS Mission.

Joint Standards. According to the *Standards for Educational and Psychological Testing* (AERA, APA, 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] and to promote valid score interpretations." The *GFTC* helps ETS to comply with the joint *Standards* because increasing fairness necessarily increases validity and reduces construct-irrelevant sources of score differences.

ETS Standards. The ETS Standards for Quality and Fairness 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 GFTC.

Acceptability of Materials. It is important to avoid irrelevant content that is commonly believed to be unfair in ETS tests or other communications. The inclusion of unnecessary content that appears to be

offensive, upsetting, controversial, or the like may lower confidence in the tests or other communications produced by ETS.

APPLICATIONS OF THE GUIDELINES

Application to Groups. Ideally, the *GFTC* applies to all relevant groups of people.[3] Special attention should be paid, however, to groups that are discriminated against on the basis of characteristics such as:

- age,
- atypical appearance,
- citizenship status,
- ethnicity,
- gender (including gender identity or gender representation),
- mental or physical disability,
- national or regional origin,
- native language,
- race,
- religion,
- sexual orientation, or
- socioeconomic status.

Application to Materials for Different Countries. The *GFTC* applies as written to materials designed primarily for use in the United States and to materials that are used worldwide. Materials designed specifically for use in a particular country other than the United States, however, will very likely require modifications to one or more of the guidelines. The needed modifications will vary depending on the country or area for which the materials are designed. For example, the guidelines for what is

[3] Not all possible groups of people are relevant in the context of fairness. For example, people born on an even-numbered day of the month and people born on an odd-numbered day are not relevant groups in the context of fairness.

considered offensive are likely to be different for tests designed for use in Qatar than for tests designed for use in Japan. Similarly, tests designed for use in Europe and tests designed for use in Africa would have different guidelines for the representation of diverse people. See *ETS International Principles for Fairness Review of Assessments* (ETS, 2009) for information on how to adapt the *GFTC* for use with materials made for particular countries other than the United States.

Application at the Design Stage. Concern with fairness in assessment begins as tests are being designed. When a construct can be measured in different ways that are comparably valid and practical, test designers should consider these guidelines in determining how best to measure the construct. For example, if a construct-relevant KSA could be measured equally well with or without the use of graphs, decisions about the appropriate way to measure the KSA should take into account the fact that graphs may cause accessibility problems for people who are visually impaired.

Tests and communications to be delivered digitally should be designed to be in compliance with guidelines for accessibility such as *Web Content Accessibility Guidelines 2.0* (World Wide Web Consortium, 2008).

Application to Tests for Different Clients. Most of the individual guidelines regarding allowable test content are based on reports about how test takers, educators, or community members have reacted to particular test content; on the beliefs of ETS staff and external consultants about what is unfair; on the preferences of clients; and on theories about the potential effects of various types of test content on test takers. Empirical support based on sound research is lacking for most of the guidelines.[4]

Different clients may reasonably have different opinions about what is considered fair. For example, one client may believe that references to social dancing are acceptable and another client may prefer to avoid the topic. Furthermore, clients may have different priorities for various aspects of test content. One client may decide, for example, that the use of unrevised, authentic stimulus materials should

[4] Much of the empirical basis for fairness guidelines is derived from studies of differential item functioning (see e.g., Holland & Wainer, 1993).

take priority over the avoidance of the use of "he" to refer to all people. Another client may prefer to avoid the use of generic "he" even if it means that the stimuli have to be revised and are less authentic.

It is appropriate for ETS to share its opinions about fairness with its clients. If, however, a client disagrees with some aspects of the *GFTC*, the fairness requirements of a client should be followed unless they would result in a violation of fairness that ETS could not accept in a test made under its auspices. For example, ETS would not allow construct-irrelevant material that incites hatred or contempt for people on the basis of disability, gender, race, religion, sexual orientation, or similar factors.

INTERPRETING THE GUIDELINES

This document contains guidelines, not rules. For many of the guidelines, compliance is a matter of degree rather than a clear binary decision. At what point does the difficulty of language become a construct-irrelevant barrier to success? How controversial does material have to be to violate a guideline? Material that seems acceptable to some reviewers may be rejected by other reviewers. 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.

An overly lax interpretation of the guidelines may allow unfair content into ETS tests. On the other hand, an overly rigid application of the guidelines may interfere with validity and authenticity. Therefore, the individual guidelines must be applied conscientiously but with regard to the client's requirements and with an awareness of the need to measure important aspects of the intended construct with realistic material that is appropriate for the test-taking population. Consider the following factors in deciding whether material is in compliance with the guidelines.

Importance for Validity. 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. In such cases, professional judgment is required to evaluate the

importance of the material for valid measurement against the extent to which the material may act as an unfair barrier to the success of some test takers.

For example, some controversial material may be important for validity. If the ability to compare and contrast two points of view about a topic is required, the topic must be controversial enough to allow at least two defensible points of view. Some offensive or upsetting material may be important in certain content areas. A history test, for example, may appropriately include material that would otherwise be out of compliance with the *GFTC* to illustrate certain attitudes commonly held in the past. Much literature deals with themes that, if not important for valid measurement, would be out of compliance with the guidelines. If controversial or offensive material is included in a test because it is important for valid measurement, ETS or the client may wish to indicate that it does not endorse the views expressed.

Type of Test or Communication. The guidelines should be interpreted differently for different types of tests, and they should be interpreted differently for tests than for communications. Skills tests assess a general skill, such as reading comprehension, writing, mathematical reasoning, or problem solving, which can be applied across subject-matter areas. Content tests assess specific subject-matter areas, such as biology, English literature, history, nursing, or psychology. Noncognitive tests measure attitudes, feelings, beliefs, interests, personality traits, and the like.

Skills tests rarely need to include material on a specific topic for valid measurement. Therefore, skills tests rarely need any material for valid measurement that is out of compliance with the *GFTC*. Content tests and noncognitive tests, however, have to include material on specific topics for valid measurement. For example, the fecal contamination of food is a topic that is too offensive to be included in a skills test of reading comprehension. The topic, however, may be important for validity in a content test for food safety inspectors, and therefore fair in that test. A noncognitive test may include, for example, questions about the test taker's attitudes toward various ethnic and racial groups. Such questions would be inappropriate in a skills test or a content test, but they might be important for validity in a noncognitive test.

It is possible that unfair material in a test may lower the scores of some test takers. It is much less likely that unfair material in a communication will directly harm the audience. Therefore, the guidelines should be interpreted more strictly for tests than for communications.

Age, Sophistication, and Previous Experience of Test Takers. In general, the older and more sophisticated the test takers, the more liberally the guidelines should be interpreted. Also, 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. Brief prior exposure does not necessarily justify the inclusion of upsetting material in a test. Furthermore, not everything that is discussed in class with the guidance and support of a teacher and the opportunity for students to ask questions is necessarily appropriate in a test. Unlike content discussed in class, content in a test may be a construct- irrelevant barrier to success and may possibly have negative consequences for the test taker. If, however, 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.

Control Over the Material. ETS has much more control over the material that it writes than it has over previously published material. Therefore, the guidelines should be interpreted more strictly for original ETS material used as stimuli than for material from other sources. Clients may require the use of unedited, authentic materials as stimuli 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 *GFTC*. The use of such materials is generally judged to be acceptable if the use of unrevised authentic materials is an important aspect of the intended construct or is required by the client and an effort has been made to obtain materials that minimize departures from the *GFTC*.

Directness of the Material. Items and stimuli about innocuous topics are generally acceptable, even if a scenario could be constructed in which they might possibly be upsetting for some test takers who had undergone a particular experience. 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, a mathematics 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.

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 a graphic description of how the cat caught, toyed with, and eventually clawed apart a baby robin would probably not be acceptable.

GENERAL PRINCIPLES FOR FAIRNESS

Though it is possible for reasonable people to disagree about a particular guideline, there are general principles for fairness that appear to be indisputable if fair measurement is a goal.

- Measure the important aspects of the intended construct.
- Provide scores that are valid for different groups in the intended population of test takers.
- Treat all test takers respectfully and impartially.
- Avoid construct-irrelevant barriers to the success of test takers, including those with disabilities and English-language learners.[5]
- Avoid construct-irrelevant content that raises strong negative feelings in test takers or others.

The guidelines that follow are intended to support these general principles.[6] If there is a disagreement about the interpretation or application of a guideline, follow the interpretation or application that best supports the general principles for fairness in assessment.

[5] References to English-language learners in the *GFTC* apply generally to learners of the language of the test, as long as the language is construct irrelevant.

[6] Even though the following guidelines focus primarily on tests, the relevant aspects of the guidelines should be generalized to apply to all communications.

GUIDELINES REGARDING CONSTRUCT-IRRELEVANT COGNITIVE BARRIERS

Construct-irrelevant cognitive barriers to success may arise when knowledge that is not part of the construct is 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. If construct- irrelevant knowledge or skill is required to answer an item and the knowledge or skill is not equally distributed across groups, then the fairness of the item is diminished.

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 other groups of test takers are, the item would be less fair. 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, fair. Whether a particular KSA is important for valid measurement or is a source of construct-irrelevant differences depends on the intended construct.

Language. Use the most accessible level of language that is consistent with valid measurement. While the use of accessible language is particularly important for test takers who have limited English skills, the use of such language is beneficial for all test takers when linguistic competence is construct irrelevant. Appendix 1 provides information about the use of accessible language in tests.

Avoid requiring knowledge of specialized vocabulary 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 in deciding which words are too specialized. Even if it is not necessary to know a construct-irrelevant, difficult word to answer an item correctly, the word may intimidate test takers or otherwise divert them from responding to the item. Similarly, avoid requiring construct-irrelevant knowledge of idioms (e.g., “flash in the pan,” “fly in the ointment”) to answer an item correctly, unless the meaning is made clear by context.

Difficult words and language structures may, of course, be used 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 word is defined in the test or its meaning is made clear by context. Complicated language structures may be appropriate if the purpose of the test is to measure the ability to read challenging material.

Specialized Knowledge. Avoid requiring construct-irrelevant, specialized knowledge to answer an item correctly. For example, knowing the number of players on a soccer team would be construct relevant on a licensing test for physical education teachers but not on a mathematics test.

Obviously, 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 expected of intended test takers are acceptable, even if the aspects are construct irrelevant. Do not, however, require specialized knowledge of these subjects unless it is construct relevant. The subject areas that require extra caution regarding specialized knowledge include, but are not limited to

- agriculture,
- construction,
- finance,
- fine arts,
- law,
- medical topics,
- military topics,
- politics,

- science,
- sports,
- technology,
- tools, and
- transportation.

For example, even if the test takers are adults, do not assume that almost all will know about a combine, a joist, a margin call, an aria, a subpoena, a stenosis, an RPG, a filibuster, a lumen, a bunt, a buffer, a chuck, or a sloop. At the appropriate grade levels, almost all are likely to know about a tractor, a board, a bank, a song, a judge, a skull, a gun, a senator, a thermometer, a ball, a computer, a hammer, or a sailboat.

Contexts. In skills tests, stimuli, such as reading-comprehension passages, have to be about something. Similarly, applications of mathematics usually require some real-world setting. The contents 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. Is it fair to have a reading passage about snow when some students have never experienced it? Is it fair to set a math problem about calculating a perimeter in the context of a fence around a garden when some students have never had a garden? What construct-irrelevant contexts are fair to include in tests? In short, the answer depends on what test takers in a particular grade 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 in a particular grade than are their home or community-based experiences.

A very important purpose for reading is to learn new things. It could severely diminish validity to limit the contents 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 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 the relevant grades are a very helpful source of information about what is considered common knowledge at those grades.

For test takers with disabilities, there is an additional requirement that direct, personal experience unavailable to the disabled test takers not be required to understand the context. For example, a test taker who is unable to participate in a footrace can still understand a problem set in the context of a footrace. On the other hand, a passage about the emotional impact of colors may be inappropriate for test takers who have never been able to experience colors. Therefore, it is best to avoid construct-irrelevant contexts that require direct, personal experience to be understood, if those experiences are not available to people with certain disabilities.

Regionalisms. Do not require knowledge of words, phrases, and concepts more likely to be known by people in some regions of the United States than in others 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. Knowledge of their meaning should not be required 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.

Religion. Do not require construct-irrelevant knowledge about any religion to answer an item. If the knowledge is part of the construct, take care to 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.

United States 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 the United States 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 quantitative reasoning.

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

- brands of products,
- celebrities,
- corporations,
- culture,
- customs,
- elections,
- food,
- geography,
- government agencies,
- history,
- holidays,
- institutions,
- laws,
- measurement systems (degrees Fahrenheit, inches, pounds, quarts, etc.),
- money,
- organizations,
- places,
- plants,
- politicians,

- political subdivisions (local, state, federal),
- political parties,
- political systems,
- public figures,
- slang,
- sports,
- sports figures,
- television shows, or
- 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” without explanation 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 immodestly. 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), the victory sign (first two fingers extended and spread in a V, other fingers clenched), and the thumbs-up sign.

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.

GUIDELINES REGARDING CONSTRUCT-IRRELEVANT AFFECTIVE BARRIERS

Construct-irrelevant affective barriers to success arise if language 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. No group of test takers should have to face language or images that are unnecessarily contemptuous, derogatory, exclusionary, insulting, or the like. 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.

Any topic that is important for validity, and for which there is no similarly important substitute, may be tested. Such topics, however, must be treated in as balanced, sensitive, and objective a manner as is consistent with valid measurement. Some stimuli and some items may necessarily focus on problematic issues. Present such material in a way that will reduce its emotional impact.

Any list of troublesome topics can be only illustrative rather than exhaustive. Current events, such as a highly publicized terrorist attack or a destructive natural disaster, can cause new topics to become distressing at any time. A topic is not necessarily acceptable merely because it has not been included in the following discussion of troublesome topics. Therefore, it is a good practice to obtain a fairness review of any potentially problematic material before time is spent developing it.

Accidents, Illnesses, or Natural Disasters. Avoid unnecessarily dwelling on gruesome, horrible, or shocking aspects of accidents, illnesses, or natural disasters. Other aspects of those topics may be acceptable. For example, it is acceptable to address the prevention of accidents, the causes of illness, or the occurrences of natural disasters. For some content tests, such as a licensing test for nurses, details about the effects of diseases or detailed descriptions of injuries may be appropriate.

Advocacy. Items and stimulus material should be neutral and balanced whenever possible. Do not use test content to advocate for any particular 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 tests may advocate for those causes on which ETS has taken a position.

Biographical Passages. It is generally best to avoid passages that focus on individuals who are readily associated with offensive or controversial topics, unless important for valid measurement. It is prudent to avoid biographical passages that focus on celebrities who are still living; their future actions are unpredictable and may result in fairness problems.

Brand Names. It is best to 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. Communications other than tests 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 groups of test takers who may favor the other side.

Highly Controversial Topics. As noted above, no topic is completely excluded from ETS tests if it is important for valid measurement and no comparably important substitute is available. Some topics, however, are so controversial, so inflammatory, or raise such strong negative emotions that they are best avoided as the focus of test materials whenever possible. If they must be included, take great care when dealing with topics such as

- abortion,

- abuse of people or animals,
- atrocities or genocide,
- contraception,
- euthanasia,
- immigration,
- painful or harmful experimentation on human beings or animals,
- hunting or trapping for sport,
- rape,
- Satanism,
- torture, and
- witchcraft.

Cryptic References. Materials used in tests come from many sources. Some of those sources may contain cryptic references to drugs, sex, racism, and other inappropriate topics. Be alert for such references and try to avoid them in tests unless they are construct relevant. Cryptic references can be a problem because test developers may be unaware of their meanings. Check the meanings of words, names, numbers, or images that appear to be arbitrary, out of place, or strange.

Some cryptic references substitute numbers for letters (1 = A, 2 = B, etc.). For example, the number 311 (three times K, the eleventh letter) has been used to stand for "Ku Klux Klan." Other cryptic numbers come from various sources. For example, the number 666 is associated with Satanism, the date April 20 is Hitler's birthday, and the time 4:20 has 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" and the word "rahowa" stands for "racial holy war."

Death and Dying. Do not focus on gruesome details associated with death and dying unless important for valid measurement. A statement that someone died in a particular year or that a disease was responsible for a certain number of deaths is acceptable.

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, for skills tests, avoid items or stimuli concerning the evolution of human beings and the similarities of human beings to other primates. Other aspects of evolution, such as the evolution of antibiotic-resistant strains of bacteria, are acceptable in skills tests. Any aspect of evolution is allowed on content tests 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 fossils or the age of Earth. Please see *Additional Guidelines for Fairness of NAEP and K–12 Tests* below for more information.

Gender. Do not assume that a pair or even a larger set of discrete categories necessarily includes the genders of all people. In particular, do not base the answer to an item on the unstated assumption that the category "male" plus the category "female" always includes all people. Consider the following item: "There are 20 adults in a room. If 8 are men, how many are women?" It is impossible to answer the question because the intended key depends on the unjustified assumption that all of the adults are either men or women. Similarly, do not assume that a married couple necessarily consists of a man and a woman.

ETS recommends asking test takers to identify their gender only if the data are required for an important purpose, such as studies of differential item functioning or reporting average scores by group. ETS also recommends adding the options "Other" and "Prefer not to respond" to the traditional "Male" and "Female" options if the client agrees to do so and the test takers are sufficiently mature.

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, and quality of culture.

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 majority culture is superior and that any differences from it constitute deprivation.

Humor, Irony, and Satire. Treat humor carefully because people may not understand the joke or may be offended by it. Similarly, treat irony and satire very carefully. In particular, avoid construct-irrelevant humor 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 some literature tests.

Images. Unless they are construct-relevant, avoid images that depict content described in other guidelines as material to avoid. Avoid construct-irrelevant images of objects or actions that have become controversial or offensive themselves (e.g., the Confederate flag, a burning cross, the Nazi salute).

Luxuries. Avoid depicting situations that are associated with excessive spending on what members of the test-taking population would consider luxuries, unless the depiction is important for validity.

Personal Questions. Unless important for validity, avoid asking test takers to respond to excessively personal questions regarding themselves, their family members, or their friends.[8] Questions about topics such as the following are generally considered inappropriate:

- antisocial, criminal, or demeaning behavior,
- family or personal wealth (unless required to determine qualification for some program or benefit),
- political party membership,
- psychological problems,
- religious beliefs or practices or membership in religious organizations, and
- sexual practices or fantasies.

[7] This is not intended to prohibit the use of reference groups in statistical analyses.
[8] In some cases there may be a need to obtain the approval of the ETS Committee on Prior Review of Research Involving Human Subjects before asking about such topics.

Profanity. Avoid blasphemy, expletives that are commonly deleted, obscenity, profanity, swear words, and the like unless important for valid measurement in literary or historical material for relatively mature test takers.

Religion. It is safest to avoid material that focuses on any religion, any religious group, any religious holidays, any religious practices, any religious beliefs, or anything closely associated with religion (including the creation stories of various cultures) unless it is important for valid measurement.

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 in ETS tests. Do not praise or ridicule the practices of any religion. Try to avoid 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, and "crusade" or "crusader" outside of their historical context). 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 proselytizing. Be factually correct and neutral in any mention of 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.

Sexual Behavior. Avoid explicit descriptions of human sexual acts unless important for validity in content tests, such as those for medical personnel. Avoid double entendres or sexual innuendo unless important for validity, such as in literature tests for relatively mature test takers.

Slavery. Except when important for valid measurement in content tests, slavery should not be the primary focus of any material. The topic is acceptable in skills tests if the emphasis of the material is on something else. For example, a passage that focuses on the accomplishments of Frederick Douglass or a passage that focuses on the abolitionist movement would be acceptable. A discussion of the details of how slaves were packed in the holds of ships for transportation from Africa would not be acceptable in a skills test, but the same discussion might be acceptable if it were important for validity in a history test.

Though "slave" is still an acceptable term, "enslaved person" is preferred by some. "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. 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 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.

Substance Abuse. Avoid a focus on the details of substance abuse, including alcohol, tobacco, and prescription as well as illegal drugs, unless important for valid measurement.

Suicide or Other Self-Destructive Behavior. It is acceptable to mention that a person committed suicide, but it is not acceptable to focus unnecessarily on various means of suicide or other self-destructive actions, unless important for valid measurement.

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 or that all people over the age of 65 are retired. 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 the test taker 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 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, even in skills tests. 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.

GUIDELINES REGARDING CONSTRUCT-IRRELEVANT PHYSICAL BARRIERS

Types of Physical Barriers. Whenever possible, tests and related materials should be created in digital formats that are accessible to individuals with disabilities.[9] Proper coding of test items enables access with audio, refreshable braille and enlarged font. Even if applicable accessibility standards have been followed, however, some physical barriers to success may remain and some test takers with disabilities may still need accommodations.[10]

Construct-irrelevant physical barriers to success occur if aspects of tests not important for validity interfere with the test takers' ability to attend to, see, hear, or otherwise sense the items or stimuli and/or to enter a response to the item. 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 based on the diagram. Test takers with hand injuries may be unable to use an answer sheet or manipulate a computer-input device. Whenever possible, appropriate accommodations should be available if physical barriers to success occur for test takers with construct-irrelevant disabilities.

Essential aspects. Some physical aspects of various item types are essential to measure the intended construct. They are, therefore, acceptable even if they cause difficulty for some test takers, including people with disabilities. For example, to measure a test taker's ability to understand speech, it is essential to use spoken language as a stimulus, even if that spoken language is a physical barrier for test takers who are deaf or hard of hearing. Essential aspects of items are those that are important for valid measurement. They must be retained, even if they act as physical barriers for some test takers.

Helpful aspects. Some physical aspects of various item types are helpful for measuring the intended construct, even if they may cause difficulty for people with disabilities. For example, drawings are often used as stimuli to elicit writing or speech in tests of English as a second language, even though the drawings are physical barriers for test takers who are blind. Stimuli other than drawings could be used

[9] For an overview of digital accessibility issues, see Hakkinen, 2015. Material on Web sites and in computer-based delivery platforms should comply with legal requirements for accessibility. Consult the office of the ETS General Counsel for the latest requirements.

[10] Consult the ETS Office of Disability Policy for advice regarding accommodations.

in this case, so the drawings are not essential. The drawings, however, are helpful as stimuli when it cannot be assumed that the test takers share a common native language. Many technologically enhanced item types are helpful for measuring the intended construct, but some enhanced item types are likely to cause difficulty for people with certain disabilities. The benefits of using the item types must be carefully weighed against their potential for raising construct-irrelevant barriers for people with disabilities.

Fairness concerns about the helpful aspects of item types should be raised at the design stage of test development when the item type is being considered. If decisions about an entire class of items have been made and reviewed for fairness at the design stage of test development, it is not appropriate to raise fairness challenges about the class of items in later reviews of individual items. It is, of course, appropriate to raise fairness challenges for problems specific to particular items.

Unnecessary aspects. Avoid unnecessary physical barriers in items and stimuli. Some physical barriers are simply not necessary. They are not essential to measure 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 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.

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 neither essential nor helpful for measuring the intended construct, avoid them in items and stimuli:

- construct-irrelevant use of a computer mouse for visually intensive tasks such as dragging and dropping text;
- construct-irrelevant charts, maps, graphs, and other visual stimuli;
- construct-irrelevant drawings of three-dimensional solids, 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;
- visual stimuli (e.g., charts, diagrams, graphs, maps) that are more complex, cluttered, or crowded than necessary;
- visual stimuli in the middle of paragraphs;
- 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);
- fine distinctions of shading or color to mark important differences in the same 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;
- labels in a stimulus that overlap with the letters used for options in the multiple-choice items based on the stimulus;
- 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/stimulus;[11]
- 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);
- special symbols or non-English alphabets (unless that is standard notation in the tested subject, such as Σ in statistical notation); and

[11] This does not apply to the traditional option labels (A–E) for multiple-choice items, even though B and D sound alike.

- 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, such as uppercase letters for variables and lowercase letters for values of those variables in statistical notation).

All of the above are acceptable if they are essential for valid measurement or have been judged to be sufficiently helpful for valid measurement to justify their use. Avoid them if they are unnecessary or not sufficiently helpful.

In addition, ensure that audio presentations are clear enough to avoid having the quality of the audio serve as a source of construct-irrelevant difficulty. Similarly, text and images displayed on a computer screen should be clear enough to avoid having the quality of the display 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.

GUIDELINES REGARDING APPROPRIATE TERMINOLOGY FOR GROUPS

If group identification is necessary, it is generally most appropriate to use the terminology that group members prefer. Unless very important for valid measurement in historical or literary material, do not use derogatory names for groups.

ETS recommends asking test takers to identify their race or ethnicity only if the data are to be used for an important purpose, such as studies of differential item functioning or reporting average scores by group. ETS also recommends allowing test takers to select more than one response in items that ask test takers to identify their race or ethnicity.

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 once. Note that terms such as “African American” and “Native American” are not hyphenated.

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. It is not, however, appropriate to use derogatory terms, even if a client should state such a preference. It may be acceptable for material that is clearly quoted from external sources to use terms other than those listed here. In authentic historical and literary material, some violations of the guidelines may be inevitable. Such material may be acceptable when it is construct-relevant, but offensive and inflammatory terms should be avoided unless they are important for valid measurement.

People Who Are African American. The terms "Black" and "African American" are both acceptable. Note that "Black" should begin with an uppercase letter when referring to people. The terms "Negro" and "Colored" are not acceptable except when embedded in literary or historical contexts or in the names of organizations. 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.

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

People Who Are Bisexual, Gay, Lesbian, or Transgendered. Identify people by sexual orientation only when it is relevant to the construct to do so. The words "bisexual," "gay," "lesbian," and "transgendered" are all acceptable. Many other terms (e.g., "cisgender," "genderless," "gynesexual," "intersex," "pansexual," "transsexual," "two-spirit") are in use, but it is not yet clear which will become widely accepted. For a specific individual, use the term preferred by the individual, if it is known. Use the pronoun preferred by the person, if it is known. If the preferred term or pronoun is rarely used, it may be necessary to explain its meaning.

Avoid using the term "homosexual" outside of a scientific, literary, or historical context. Do not use the term "queer" to refer to sexual orientation except in reference to the academic fields of queer

theory and queer studies in institutions that use those labels. Avoid using the phrase "sexual preference" for sexual orientation. Do not refer to heterosexual relationships as "normal" and other types of relationships as "abnormal."

People with Disabilities. To avoid giving the impression that people are defined by their disabilities, focus on the person rather than the disability in the first reference to someone with a disability. The generally preferred usage is to put the person first and the disabling condition after the noun, as in "a person who is blind."

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," "crippled," "confined," "inflicted," "pitiful," "victim," "suffering from" 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," and the use of such words and phrases as "inspirational," "courageous," 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.

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 doctor is the topic. If a person is in treatment with a nonmedical professional (e.g., social worker, psychologist), "client" is the appropriate term.

Content 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.

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.

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."

People With a Cognitive Disability. Do not use the word "retarded" except when necessary in literary or historical materials. Preferable terms are "cognitively disabled," "cognitively impaired," "developmentally delayed," "developmentally disabled," or "intellectually disabled." Use the term "Down syndrome" rather than "Down's syndrome." Avoid the obsolete and inappropriate term "Mongoloid."

People With a Motor Disability. The words "paraplegic," and "quadriplegic" are acceptable as adjectives, not as nouns. The word "spastic" is unacceptable to describe a person. Muscles are spastic, not people.

People of Different Genders. When gender is not construct relevant, the general goal is to treat people as equally as possible regardless of their gender. Though the concepts of male and female do not necessarily include the genders of all people, the conventions regarding language use discussed below focus on those two genders.

Before about 1970, it was generally considered correct to refer to all human beings as "man" and to use words such as "chairman," "mankind," or "manpower" based on that convention. Therefore, it is very difficult to find authentic materials written before then that are in compliance with the following 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.

When possible, women and men should be referred to in parallel terms. When women and men are mentioned together, both should be indicated by their full names, by first or last name only, or by title. Do not, for example, indicate men by title and women by first name. The term "ladies" should be used for women only when men are being referred to as "gentlemen." Similarly, when women and men are mentioned together, 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 woman is known to prefer it. Women should not be described by physical attributes when men are described by mental attributes or professional position, or vice versa. Gratuitous references to appearance or attractiveness are not acceptable except in literary or historical material. Except in literary or historical material, women eighteen or older should be referred to as "women," not "girls." Men eighteen or older should be referred to as "men," not "boys."

Using "he" or "man" to refer to all people is not appropriate unless it is included in historical or literary material. Minimize the use of words that indicate all members of a profession or all people serving in a particular role are male (e.g., use "police officer" rather than "policeman," "supervisor" rather than "foreman"). If the gender of a subject is not specified, use "he or she" or "she or he" as pronouns (e.g., "If a student studies, he or she will pass"). Because gender is not limited to male and female, it may be preferable to use plural constructions to avoid gender-specific pronouns when the gender of the subject is not specified (e.g., "If students study, they will pass"). Alternating generic "he" and generic "she" is inappropriate because neither word should be used to refer to all people. Try to avoid the constructions "he/she" and "(s)he."

Because generic terms such as "doctor," "nurse," "poet," and "scientist" include both men and women, modified titles such as "poetess," "woman doctor," or "male nurse" are not appropriate.[12] Do not use expressions such as "the soldiers and their wives" that assume only men fill certain roles, unless such is the case in some particular instance. Do not couple generic role words with gender-specific pronouns

[12] "Actress" is acceptable when paired with "actor," as in "awards for te best actor and actress."

unless a particular person is being referenced. Do not, for example, use terminology that assumes 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.

People Who Are Hispanic American. The terms "Latino American" (for men and mixed gender groups), "Latina American" (for women), and "Hispanic American" are acceptable and may be used as appropriate. Though "Chicano" and "Chicana" as terms for Mexican Americans are accepted by some groups, they are rejected by others. It is therefore best to avoid using them. Where possible, use a specific group name such as "Cuban American," "Dominican American," or "Mexican American" as appropriate.

People Who Migrate to the USA. "Immigrant" and "migrant" are acceptable terms. For people who enter the United States illegally, use "undocumented immigrant" rather than "illegal alien." Do not use "illegal" as a noun to refer to an undocumented immigrant.

People Who Are Members of Minority Groups. Members of so-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.

The terms "biracial" and "multiracial," as appropriate, are acceptable for people who identify themselves as belonging to more than one race or ethnicity. "People of color" is acceptable for biracial and multiracial people and for 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.

People Who Are Native American. The terms "American Indian" and "Native American" are acceptable. 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.

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.

Some clients request the terminology preferred by constituent groups within their jurisdiction. Clients may, therefore, 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.

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. The term “English as a second language (ESL)” applies to people learning English in an English-speaking environment, whereas “English as a foreign language (EFL)” applies to people 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 (e.g., “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 the person’s lack of English proficiency. However, once “ELL” or “EL” has been used as an adjective, it is acceptable to refer to people as “ELLs” or “ELs.”

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 “elderly” as a noun. Minimize the use of euphemisms such as “senior citizens” or “seniors.” Tests in certain content

areas such as medicine may use terms such as “old-old” or “oldest-old” that are not appropriate in general usage.

People Who Are White. The terms “White” and “Caucasian” are both acceptable, but “White” is becoming the preferred term. Note that “White” should begin with an uppercase letter when referring to people. The term “European American” is preferred by some people because of its parallelism to “African American,” “Asian American,” “Native American,” and so forth. “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.” Descriptive uses of “white” as in “White House,” “white collar,” “white water,” and so forth are acceptable.

GUIDELINES REGARDING REPRESENTATION OF DIVERSITY

If a test mentions or shows people, test takers should not be made to feel alienated from the test because no members of their group are included. Therefore, the ideal test would include members of the various relevant 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 or show people.

Application. Follow the guidelines below for tests designed for use throughout the United States and 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 responsible assessment director to determine the characteristics of the test-taking population to be reflected in a test made for a particular country. Also, consult the client or responsible assessment director to determine the diversity to be reflected in a test made specifically for a particular jurisdiction within the United States, such as a consortium of states, a state, a city, or a school district.

Gender Balance. In skills tests, women and men should be represented in comparable ways. In addition to roughly balancing numbers of people of each gender, the status of the men and women shown should be 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 content tests should be appropriate to the subject matter. For example, most of the people mentioned in a test of military history would be men. The gender balance of occupational tests should be roughly appropriate to the gender distribution of members of the occupation.[13] For example, most of the people mentioned in a test for licensing nurses would be women and most of the people mentioned in a test for licensing automobile mechanics would be men. Strive for some diversity when possible, however.

Racial and Ethnic Balance. Because about one-third of the people in the United States are members of so-called minority groups, try to have about one-third of the items that mention people in skills tests represent people from what are commonly considered minority groups in the United States or people from the countries of origin of those groups. 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 the names of reasonably well-known real people in various groups or by using generic names commonly associated with various groups. Do not add unnecessarily to the linguistic loading of an item by using names that are inordinately difficult for test takers to decode (e.g., use "Vijay" rather than "Visalakshi").

In some content tests, such as history tests or literature tests, the proportion of items dealing with diverse groups may be indicated by the test specifications. If the proportions are not fixed by the test specifications, try to meet the representational goals given for skills tests to the extent allowed by the subject matter. If the names of people appearing in content tests are part of the subject matter (e.g.,

[13] General impressions of the gender balance of various occupations and subject areas are sufficient. Research studies to obtain more precise information are not necessary.

Avogadro's number, Heimlich maneuver, the Jay Treaty), the items are not counted as mentioning people for the purpose of calculating the number of items in which diversity should be represented.

People Who Are Bisexual, Gay, Lesbian, or Transgendered. People may be identified as bisexual, gay, lesbian, or transgendered in tests when it is construct relevant to do so. The guidelines are in conflict, however, concerning the construct-irrelevant identification of people as bisexual, gay, lesbian, or transgendered for the purpose of representing diversity. The guideline about representing diversity would argue for their representation. Other guidelines, including the restriction on identifying people by sexual orientation unless it is relevant to the construct to do so, would argue against their representation. In light of the conflict among the guidelines, identify people as bisexual, gay, lesbian, or transgendered in tests for purposes of representing diversity only with the approval of the client or governing board for the test, and only for tests administered to relatively mature test takers.

People with Disabilities. Occasionally represent people with disabilities in tests that include people. Be careful not to reinforce stereotypes when doing so, however. For example, a picture of a person in a wheelchair in a work setting may be appropriate. If, however, the person is shown being pushed by someone else, that 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.

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.[14] 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, in a skills test do not portray all of the executives as male and all of the support staff as female. Do not overrepresent members of any group in examples of inappropriate, foolish, unethical, or criminal behavior.

[14] If tests are made specifically for a particular country other than the United States, it may be necessary to modify these guidelines as described in *ETS International Principles for Fairness Review of Assessments* (2009).

ADDITIONAL GUIDELINES FOR FAIRNESS OF NAEP AND K–12 TESTS

These guidelines for NAEP and K-12 tests are in addition to the guidelines that apply to all ETS tests.

Requirements for NAEP. The following requirements are excerpted from the National Assessment Governing Board Policy Statement, *NAEP Item Development and Review,* adopted May 18, 2002.

Secular, Neutral, Non-Ideological. 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," or "Hinduism is the predominant religion in India."

Neutral and Non-ideological - 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 non-ideological also pertain to decisions about the pool of test questions in a subject area taken as a whole. The 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 non- ideological, 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)

In addition, NAEP follows the guidelines for K-12 assessments below.

K-12 Assessments. The fairness guidelines imposed by consortiums of states, individual states, cities, or school districts for their K–12 tests are often more rigorous and extensive than the requirements for other tests.[15] 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. Some of these requirements may appear overly strict to test developers who are not used to the K–12 environment, but each of the restrictions discussed below has been judged to be important by one or more jurisdictions.

Different K–12 clients have different fairness requirements. Because the constraints listed below have been compiled from several jurisdictions, no single jurisdiction is likely to require them all. Check with the responsible assessment director for the specific fairness requirements of the client. In the absence of information to the contrary for a particular client, however, it is safest to follow the generic K–12 requirements listed below in addition to the fairness guidelines that apply to all ETS tests. Content that is required by a jurisdiction's content standards may be tested, even if it would otherwise be out of compliance with the following guidelines.

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**.**

[15]Nationally administered admissions, placement, or guidance tests such as SAT®, AP®, and PSAT/NMSQT® are taken by high school students but are not considered K–12 tests for the purposes of these guidelines

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 atypical physical attributes (e.g., anorexia, disfigurement, early or late physical development, obesity, small stature, stuttering);
- dissension among family members or between students and teachers;
- serious illnesses (e.g., cancer, AIDS, 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;
- family situations that students may find upsetting (e.g., deportation, divorce, eviction, homelessness, loss of job, layoff, incarceration, separation);
- violence or conflict, including domestic violence, playground arguments, fights among students, bullying, 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);
- animals that may be sensitive topics for specific cultural groups (e.g., the owl for some Native American nations). Check the issue of animals that may be sensitive topics with the client or responsible assessment director.

Offensive topics. Unless they are important for validity, avoid topics that may be offensive to particular groups in a jurisdiction, such as

- drinking alcohol, smoking or chewing tobacco, using drugs (including prescription drugs in some jurisdictions);

- gambling (playing cards and dice may be used as required in math problems, but do not assume that all students will be familiar with them);
- holidays and other occasions alien to some students (e.g., birthday celebrations, Halloween, Valentine's Day);
- social dancing, including school dances (such as proms), particular kinds of music (e.g., rap, rock and roll), attending movies (these sensitivities vary greatly by client);
- references to a deity, including expressions like "thank God" and euphemisms for references 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; and
- texts that are preachy or moralistic, as 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 particular 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. Such passages or stimuli should be clearly labeled as persuasive or editorial text. Topics that are particularly troublesome in some jurisdictions include

- 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;
- labor unions;

- prayer in school;
- the suffering of individuals at the hands of a prejudiced or racist society, a focus on individuals overcoming prejudice, or the specific results of discrimination; and
- welfare or food stamps.

Inappropriate behavior. Do not use material that models or reinforces inappropriate student behaviors. In particular, 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, or running away from home, or even considering those behaviors;
- 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), even if everything turns out well;
- sexual activity, unless required by content standards; and
- expressing or implying cynicism about charity, honesty, or similar values esteemed by the community.

Specific content areas. 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 assessment, although a similar discussion should be avoided in other content areas. The subject of evolution (with associated topics of fossils, dinosaurs, and geologic ages) can be included in a K–12 test if it is required by a content standard. 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.

ADDITIONAL FAIRNESS ACTIONS

The GFTC is meant to help the people who design, develop, and review ETS items and tests make fair and valid tests, but the use of the GFTC alone is insufficient. Additional fairness actions such as the following are required as well.

Impartiality. An important aspect of fairness is treating people impartially, regardless of personal characteristics that are not relevant to the test being given (such as gender, race, ethnicity, or disability). ETS strives to give all test takers respectful treatment, equal access to relevant testing services, and useful information about the assessment. As part of treating test takers appropriately, ETS strives to create websites, information, test preparation materials and tests in digital formats accessible to persons with disabilities. In addition to meeting accessibility standards, ETS provides needed accommodations to test takers with disabilities so that the test measures relevant knowledge and skills rather than the irrelevant effects of a person's disability.

External Contributors. ETS encourages contributions to tests from external people who represent relevant perspectives and diverse groups. Representatives of various groups are included in the test-development committees that determine the knowledge, skills, and other attributes to be tested. Committee members may also write, review, revise, and select the items to be included in the test. Additional means of obtaining contributions to help maintain fairness include involving people who are members of various racial and ethnic groups as external item writers and reviewers, as test reviewers, and as essay scorers.

Differential Item Functioning (DIF). As an empirical check on the fairness of items, statistical measures of differential item functioning (DIF) are used. DIF occurs when people in different groups perform in substantially different ways on a test item, even though the people have been matched in terms of their relevant knowledge and skill as measured by the test. The statistics are applied whenever the data would be useful and sample sizes are large enough to allow meaningful results. If DIF data are available, tests are assembled following rules that keep DIF low. If data are unavailable at assembly, DIF is calculated after test administration. Items with high DIF are reviewed for fairness by panels of people

who have no vested interest in the test. Any items judged to be unfair are removed before the test is scored. See Dorans (1989) and Zieky (1993) for more information about DIF.

Validation. A crucial aspect of fairness is validation. Essentially, validation is the collection of evidence to evaluate the extent to which the inferences made on the basis of test scores are appropriate. Multiple lines of evidence are pursued in validation efforts. Some important aspects of validation are, for example, demonstrating that the people who determined the specifications for the test had the training and experience necessary to do a competent job; showing that the different parts of the test relate to one another and to external criteria as theory would predict; and determining the extent to which the items sample only relevant knowledge and skills. See Messick (1989) and Kane (2006, 2013) for more information about validity.

Test Interpretation and Use. Even a fair test can be used unfairly. For example, when the opportunity to learn the tested material is not equally distributed, interpreting scores as measures of the ability to learn is unfair. ETS specifies the appropriate interpretation and use of its tests and makes the information available to score recipients. ETS or the client investigates plausible allegations of misuse and informs the score recipient how to use the test appropriately.

Research. ETS supports a great deal of research directly related to test fairness that is too extensive to cite here. Free reports of the research are available to the public at ETS's website, www.ets.org. Click on the tab for Research and use the search facility with key words such as “African American,” “Asian,” “bias,” “Black,” “Caucasian,” “culture,” “DIF,” “differential item functioning,” “differential prediction,” “differential validity,” “disability,” “disadvantaged,” “fair,” “fairness,” “female,” “gender,” “group score differences,” “handicap,” “Hispanic,” “Latino,” “male,” “minority,” “Native American,” “Negro” (for older reports), “sensitivity,” “underprivileged,” “underrepresented,” “underserved,” “unfair,” or “White.”

CONCLUSION

The task of developing guidelines for the fairness of tests and communications 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 particular test or communication, created at a particular time, for a particular purpose, and for a particular population. If appropriately interpreted and applied, however, the GFTC will help to ensure that fairness is an important consideration as tests and communications are developed and reviewed, and will help ETS attain the goal of making tests and other communications as fair as possible.

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association.* Washington, DC: APA.

Associated Press (AP). (2015). *The Associated Press stylebook* (46th ed.). New York: Basic Books.

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

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

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

ETS. (2009). ETS international principles for fairness review of assessments. Princeton, NJ: Author.

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* (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.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (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.

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.

University of Chicago Press. (2010). *The Chicago manual of style* (16th ed.). Chicago: University of Chicago Press.

World Wide Web Consortium. (2008). Web content accessibility guidelines 2.0. Retrieved from http://www.w3.org/TR/WCAG20/

Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

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.

APPENDIX 1: ETS GUIDELINES FOR USING ACCESSIBLE LANGUAGE

The purpose of this appendix is to help those who create ETS tests make the language in ETS tests as accessible as possible to all test takers.

While some test takers may especially benefit from the use of accessible language (e.g., those with limited knowledge of English, those with disabilities related to language processing, those who are not strong readers), accessible language is not a testing accommodation. It is a practice that seeks to minimize construct-irrelevant variance for all test takers.

In some cases, clients have specific style guides or other policies related to language use. The *ETS Guidelines for Using Accessible Language* is not intended to conflict with client preferences.

Application. These guidelines apply to all test takers, to all elements of a test (directions, stimuli, stems, options, etc.), and to all associated test material (registration bulletins, 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 more accessible.

The following are some specific examples of instances where prioritizing accessible language is inappropriate:

- 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 nonconstruct-related aspects of the test material should use the most accessible level of language that is consistent with validity. In general, this means that the stimulus should be as complicated as necessary to assess the construct, while items should be as accessible as possible.

Guidelines for Accessible Language. Writing in a clear and accessible way requires diligence and a clear understanding of what you are trying to say. In striving to cast tests in the most accessible level of language possible, consider the audience and the construct being assessed and continually look for ways to increase clarity and improve comprehension. The ideas presented below should be thought of as guidelines to help meet that goal rather than as rules to be followed strictly.

Paragraphs

- Try to use short, clear paragraphs. In most text, paragraphs should contain fewer than 150 words. In expository writing, most paragraphs should have one main idea and should state it in the first or second sentence.

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 in the sentence "The book that I am reading is interesting."). While relative clauses can be an effective means of representing complex ideas in a single sentence, their overuse can decrease accessibility by making 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.
- 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" if necessary for clarity.

Vocabulary

- Use vocabulary that is widely accessible to test takers. Whenever possible, use common words rather than less common synonyms (e.g., "walk" rather than "ambulate").[16][16]
- Try to use specific, concrete words rather than more abstract words (e.g., "house" rather than "dwelling"). Note, however, that it is appropriate for a history test to ask a question such as "What type of dwelling did the Iroquois live in before European exploration?" In that case, "dwelling" is appropriate because the options may include living spaces other than houses.
- Avoid the use of foreign expressions that may be less familiar than common English equivalents (e.g., "in lieu of" versus "instead of").
- 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.
- 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., "computer modem 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).

[16] 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

- 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."

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.

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, "Mark the best answer to each question" is clearer than "Each student should mark the best answer to each question."

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.

Some Particular Issues for Test Items. The guidelines above apply to all types of writing. Below 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.

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 much 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 three 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 three 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 "NOT" in all caps) to reinforce that the stem is negative.

Contexts for Word Problems. When a context is to be provided (e.g., for a mathematics word problem), use a context that is no more complicated than required for valid measurement. Try to use contexts that will be familiar to as wide a range of test takers as possible. Remember that students who are from outside the United States, students who have economic disadvantages, or students with disabilities may not have had the same experiences as other students. At the K–12 level, a school-based context is often accessible to a wider range of students than a home-based context is.

If a mathematics construct can be assessed just as well without the use of a context, consider omitting the context. Mathematics problems are sometimes given an empty context that is irrelevant to the construct being assessed. For example: "Last month, 193,825 people visited the museum. What is the value of 8 in 193,825?" If the first sentence is deleted, the item becomes more accessible, while the mathematical task remains unchanged.

Examples. The following examples have been selected to show how the language of test items can be modified to increase accessibility 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: Removing the "if," adding "a student" to serve as the subject of an active verb, and dividing one complex sentence into two simpler sentences all help to present the task more clearly.	

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?	Vijay 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.	

APPENDIX 2: ABRIDGED LIST OF GUIDELINES FOR FAIRNESS

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

The purpose of this highly abridged listing of fairness guidelines is to help ETS staff use and refer to the guidelines in their daily work. The body of the *GFTC* includes explanations, discussions, caveats, and examples to help readers understand and interpret the guidelines appropriately. Those features help users learn about the guidelines but 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.

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

Groups of Particular Concern. Although the *GFTC* applies to all relevant groups of people, special attention should be paid to groups that have been or are discriminated against on the basis of characteristics such as

- age,
- atypical appearance,
- citizenship status,
- ethnicity,
- gender (including gender identity or gender representation),
- mental or physical disability,
- national or regional origin,
- native language,
- race,
- religion,
- sexual orientation, or

- socioeconomic status

Interpreting the Guidelines. To interpret the guidelines correctly, take into account the opinions of the client; the need for authenticity and the importance for validity of the material; the age, sophistication, and previous experiences of test takers; the degree of ETS control over the material; the directness and extent of the material; and the type of test or communication.

General Principles for Fairness.

- Measure the important aspects of the intended construct. Provide scores that are valid for different groups in the intended population of test takers.
- Treat all test takers respectfully and impartially.
- Avoid construct-irrelevant barriers to the success of test takers, including those with disabilities and English-language learners.
- Avoid construct-irrelevant content that raises strong negative feelings in test takers or others who are concerned about test materials.

A) Guidelines Regarding Construct-Irrelevant Cognitive Barriers

1. **Accessible language**. Use the most accessible language that is consistent with valid measurement.
2. **Specialized knowledge**. Avoid requiring construct-irrelevant specialized knowledge of a subject to answer an item correctly.
3. **Contexts**. For construct-irrelevant contexts of reading passages and math problems, the information required to key the items correctly should either be common knowledge among the intended test takers or be available in the passage or problem. Contexts should not require direct experience that is unavailable to people with disabilities.
4. **Regionalisms**. Do not require construct-irrelevant knowledge of words, phrases, and concepts more likely to be known by people in some regions of the United States than in others.

5. **Religion**. Do not require construct-irrelevant knowledge about any religion to answer an item.
6. **United States culture**. Do not require a test taker to have construct- irrelevant specific knowledge of the United States to answer an item. Do not assume that all test takers are from the United States.

B) Guidelines Regarding Construct-Irrelevant Affective Barriers. The word "avoid" should be interpreted as though it were followed by the words "unless important for valid measurement."

1. **Accidents, illnesses, disasters**. Avoid dwelling on gruesome, horrible, or shocking aspects of accidents, illnesses, or natural disasters.
2. **Advocacy**. Items and stimulus material should be neutral and balanced whenever possible. Do not use test content to advocate any particular cause or ideology or to take sides on any controversial issue.
3. **Biographical passages**. Avoid passages about people associated with controversial or offensive topics. It is prudent to avoid passages about live celebrities.
4. **Brands.** Avoid construct-irrelevant brand names.
5. **Conflicts.** Do not focus on conflicts in which test takers may sympathize with different factions. Do not take sides.
6. **Controversial topics**. Take great care when dealing with particularly problematic topics such as abortion, genocide, rape, or torture. It is best to avoid them if possible.
7. **Cryptic references.** Avoid cryptic references to inappropriate topics (e.g., sex, drugs, White supremacy). Check names, numbers, and words that seem strange, arbitrary, or out of place.
8. **Death.** Do not focus on gruesome details associated with death and dying.

9. **Evolution.** For skills tests, avoid items or stimuli concerning the evolution of human beings. Evolution is acceptable for content tests as necessary. Check with the client or responsible assessment director for K–12 tests.
10. **Gender.** Do not assume that "male and female" includes all people. Do not assume a married couple necessarily includes a man and a woman.
11. **Group differences**. Avoid unsupported generalizations about the existence or causes of group differences.
12. **Humor, irony, and satire**. Treat humor, irony, and satire very carefully.
13. **Images.** Avoid images depicting content that other guidelines identify as material to avoid. Avoid controversial or offensive images.
14. **Luxuries.** Avoid depicting situations that are associated with spending money on what test takers would consider luxuries.
15. **Personal questions.** Avoid asking test takers to respond to excessively personal questions regarding themselves, their family members, or their friends.
16. **Profanity.** Avoid blasphemy, obscenity, profanity, swear words, and the like.
17. **Religion.** Avoid material that focuses on any religion, any religious group, any religious holidays, any religious practices, any religious beliefs, or anything closely associated with religion (including the creation stories of various cultures).
18. **Sexual behavior.** Avoid explicit descriptions of human sexual acts. Avoid double entendres and sexual innuendo.
19. **Slavery.** Reference to slavery is acceptable in content tests if important for validity. Such references are acceptable in skills tests only if the emphasis of the material is on something else.
20. **Stereotypes.** Avoid stereotypes. If some group members are shown in traditional roles, other members of the group should be shown in nontraditional roles.

21. **Substance abuse.** Avoid focusing on the details of substance abuse, including alcohol, tobacco, prescription medications, and illegal drugs.
22. **Suicide or other self-destructive behavior.** It is acceptable to mention that a person committed suicide, but do not focus unnecessarily on various means of suicide or other self-destructive behavior.
23. **Unstated assumptions.** Avoid material based on underlying assumptions that are false or that would be inappropriate if the assumptions were stated.
24. **Violence and suffering.** Do not focus on violent actions, on the detailed effects of violence, or on suffering.

C) Guidelines Regarding Construct-Irrelevant Physical Barriers. If these barriers, or others like them, are construct irrelevant and neither essential nor helpful for measuring the intended construct, avoid

1. visual material that is unnecessary, needlessly complicated, or hard to discern;
2. letters or symbols that are easily confused with one another or are uncommon;
3. intermingling text and illustrations; and
4. printing text vertically or on a slant.

D) Guidelines Regarding Appropriate Terminology for Groups

	GROUP	**OK**	**AVOID**
1	All	Name that group members prefer	Derogatory names
2	All	Names as adjectives (e.g., "Black students")	Names as nouns (e.g., "the Blacks")
3	People Who are African American	Black and African American	Colored, Negro (except in literary and historical material or in the name of an institution)

	GROUP	OK	AVOID
4	People Who are Asian American	Asian American, Pacific Island American, and Asian/Pacific Island American (used as appropriate) Use specific group names (e.g., "Chinese American," "Japanese") when possible.	Oriental (except in literary and historical material or in the name of an institution)
5	People Who are Bisexual, Gay, Lesbian, or Transgendered	Bisexual, gay, lesbian, transgendered (or term person prefers, if known) Sexual orientation	Gratuitous labels Homosexual, queer Sexual preference
6	People With Disabilities	Put the person first and the disabling condition after the noun (e.g., "a student with . . .").	Negative terms Euphemistic terms
7	People Who are Blind	Blind Visually impaired	The blind (except in literary and historical material or in the name of an organization)
8	People Who are Deaf	Deaf Deaf and hard of hearing	Deaf and dumb, deaf mute, hearing impaired The Deaf (except in literary and historical material or in the name of an organization)
9	People With a Cognitive Disability	Cognitively impaired, developmentally disabled, developmentally delayed, or intellectually disabled	Retarded Mongoloid
10	People With a Motor Disability	The terms "paraplegic" and "quadriplegic" are acceptable as adjectives but not as nouns.	Spastic (to describe a person)

	GROUP	OK	AVOID
11	People of Different Genders	Parallel terms The phrase "he or she" Pluralization (e.g., "they") Gender-free terms (e.g., "fire fighter," "police officer")	Nonparallel terms He or man (to refer to all people) Male terms (e.g., "fireman," "policeman") Gratuitous references to appearance Girl or boy (for people 18 or over) Assumption that all members of a profession are one gender (e.g., "doctor" includes people of all genders)
12	People Who are Hispanic American	Latino American (for men or mixed- gender groups), Latina American (for women), and Hispanic American Specific group name, such as Cuban American or Cuban	Chicano, Chicana
13	Immigrants	Immigrant or migrant (for people who enter legally) Undocumented immigrant (for people who enter illegally)	Illegal alien Illegal (as a noun)
14	People Who are Multiracial	Biracial, multiracial, people of color	Colored people (except in literary and historical material or in the name of an organization)
15	People Who are Native American	American Indian, Native American, names of specific tribes or nations	Eskimo, buck, squaw, the word "brave" as a noun (except in literary and historical material)

	GROUP	OK	AVOID
16	Nonnative Speakers of English	ELL, EL, ESL, EFL (as adjectives) ELL, EL (as a noun after first use as an adjective)	ESL, EFL (as nouns)
17	People Who are Older	Specific ages or age ranges	Elderly (as noun) Euphemisms
18	People Who are White	White, Caucasian, European American	Anglo American (unless meaning is clear from context)

E) Guidelines for Representing Diversity

1. Gender balance. In skills tests, women and men should be comparably represented. The gender balance of content and occupational tests should be appropriate to the content area or occupation.
2. Racial and ethnic balance. For skills tests, about one-third of the items that mention people should represent people from what are commonly considered minority groups in the United States or people from the countries of origin of those groups. For content and occupational tests, try to meet the representational goals given for skills tests to the extent allowed by the subject matter.
3. People who are bisexual, gay, lesbian, or transgendered. Because guidelines are in conflict regarding the construct-irrelevant representation of people who are bisexual, gay, lesbian, or transgendered, represent them in tests for mature test takers, and only with the approval of the client or governing board for the test.
4. People with disabilities. Occasionally represent people with disabilities in tests that include people.
5. 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.
6. Additional Guidelines for Fairness of NAEP and K–12 Tests

7. The following requirements for NAEP and K–12 tests are in addition to the fairness guidelines that are to be followed for all ETS tests. For K–12 tests, check with the responsible assessment director for the specific fairness requirements of the client.
8. NAEP. Items should avoid asking about sensitive topics and be
9. secular (e.g., neither for nor against religion),
10. neutral (e.g., not taking sides in controversies), and
11. nonideological (e.g., not advocating for particular causes or lifestyles).
12. K-12. Avoid topics that
13. are emotionally charged for children (e.g., death, disasters, divorce, violence);
14. could be considered offensive (e.g., drinking, gambling, smoking);
15. are controversial in the jurisdiction of the test (e.g., immigration, prayer in school); or
16. encourage inappropriate behavior for children (e.g., cheating, fighting, lying).

##########

###

Measuring the Power of Learning.®

www.ets.org