Listening. Learning. Leading.®

ETS Guidelines for Fairness Review of Assessments

2009

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 fairness review process. Fairness review is an essential tool in accomplishing the ETS mission "to help advance quality and equity in education by providing fair and valid assessments."

ETS has required a documented fairness review based on published guidelines for every test it has made since 1980. Training in performing fairness reviews and strict adherence to the published guidelines have been required of all ETS test developers since that time.

The guidelines are not static. They were expanded and updated seven times between 1980 and 2003. As societal views of fairness have evolved and as knowledge of the relationship between fairness and validity has increased, each successive version of the guidelines has been more inclusive and comprehensive than the previous one.

This latest version has maintained the tradition of continual improvement. It provides a firm basis for fairness review in validity theory, expands the number of groups of concern for fairness reviewers, adds a new guideline intended to make tests more accessible for people with disabilities, and includes a new chapter on special considerations for the fairness of K–12 assessments. I am pleased to issue the 2009 version of the *ETS Guidelines for Fairness Review of Assessments.*

Glenn Schroeder
Senior Vice President and General Counsel
Educational Testing Service

INTRODUCTION

Purpose: The primary purpose of the *ETS Guidelines for Fairness Review of Assessments* is to enhance the fairness of tests. These guidelines are intended to help the people who design, develop, and review ETS items and tests to

- better understand fairness in assessment,
- avoid the inclusion of unfair content or images in tests as they are developed,
- find and eliminate any unfair content or images in tests as they are reviewed, and
- reduce subjective differences in decisions about fairness.

The guidelines are written for ETS staff and for the external subject-matter experts who help them design, develop, and review items and tests.[1] The guidelines are copyrighted but not confidential. In accordance with its mission to improve testing, ETS allows use of the guidelines by all who wish to enhance the fairness of their tests.

Overview: These guidelines provide rationales for fairness review, including the close relationship between fairness and validity. This document describes the three guidelines to be applied to all ETS tests and describes additional constraints for ETS K–12 tests. Following the descriptions are two appendices: appendix 1 contains the *ETS Guidelines for Using Accessible Language*; appendix 2 describes the actions that ETS takes in addition to fairness review to help ensure the fairness of tests. A list of references regarding fairness in assessment is also included.

Rationales for Fairness Review

ETS Mission: The ETS mission is, in part, "to help advance quality and equity in education by providing fair and valid

[1] Guidelines for ETS material other than tests are included in a separate publication, *ETS Guidelines for Fairness Review of Communications* (ETS, 2009).

assessments." Fairness review is an essential step in the provision of fair and valid assessments because fairness review helps to ensure that ETS tests show respect for diverse groups of people, are sensitive to the needs and feelings of the intended test takers, avoid images and content that are insulting or demeaning, and are free of unnecessary barriers to the success of all test takers. Fairness review helps to ensure that the people who take ETS tests are not unnecessarily distracted by being angered, irritated, or frustrated by them.

ETS Standards: The *ETS Standards for Quality and Fairness* (2002) requires ETS staff to "ensure that symbols, language, and content that are generally regarded as sexist, racist, or offensive are eliminated, except when necessary to meet the purpose of the assessment."[2] The mechanism used to meet that requirement is fairness review.

Fairness and Validity: When people use test scores, they make inferences about the knowledge, skills, or other attributes (KSAs) of test takers on the basis of the scores. The extent to which those inferences are appropriate is an important aspect of validity. The extent to which those inferences are appropriate for different groups of test takers is an important aspect of fairness. Fairness and validity are very tightly linked in assessments because a fair test is one that is valid for different groups of test takers in the intended population for the test.

Therefore, fairness review is not a mere exercise in political correctness. Fairness is essential for valid measurement, and validity is essential for fair measurement. Practices that increase fairness will increase validity, and practices that increase validity will increase fairness.

Nothing in these guidelines is intended to interfere with valid measurement. Material that is important for valid measurement—and for which a more appropriate substitute

[2] For example, to meet its purpose, a licensing test for physicians may require material that would be considered offensive in a test of reading comprehension for high school students.

is not available—is acceptable, even if it includes content or images that these guidelines would otherwise prohibit.

Constructs and Variance: In the context of testing, the construct is defined as all of the KSAs that a test, or a section of a test, is supposed to measure. Variance is a label for differences among test scores.[3] If everybody gets the same score, the variance is zero. As scores become more and more spread out, the variance increases.

In an ideal test, all of the variance would be caused by construct-relevant differences in the KSAs of the test takers.[4] Sometimes, however, the construct that is supposed to be measured by a test is contaminated by other sources of variance that the test may inadvertently include. For example, if the language in the items is unnecessarily difficult, a test intended to measure quantitative reasoning may inadvertently include verbal ability as a source of variance.

Construct-Irrelevant Variance: Any KSAs or other sources of variance that are not in the intended construct are called construct-irrelevant sources of variance. Any construct-irrelevant sources of variance that would cause test takers who actually have the required KSAs to answer an item incorrectly (or cause test takers who actually lack the required KSAs to answer an item correctly) will lead to incorrect inferences about the test takers and, therefore, diminish validity.

If certain groups of test takers (e.g., English language learners, African American students, female job applicants, people who have impaired vision) are significantly more affected by a construct-irrelevant source of variance than are other groups, then fairness is diminished as well as validity. Fairness review en-

[3] Variance is also a statistical term, but knowledge of statistics is not required to understand these guidelines.

[4] No test meets the ideal, because there are always some elements of random chance in the score variance. For example, a test taker may make a lucky guess on a multiple-choice item, or a test taker's essay may happen to be scored by the most lenient of the readers, thus causing some random score variance.

hances the fairness and validity of assessments by removing identifiable construct-irrelevant sources of variance that may affect different groups in different ways.

Rationale for the Guidelines

Sources of Variance: Sources of construct-irrelevant variance can be cognitive, affective, or physical.[5] The three sources of construct-irrelevant variance lead directly to the three guidelines for fairness review of ETS assessments.

1. Avoid cognitive sources of construct-irrelevant variance.

   Cognitive sources of construct-irrelevant variance occur when knowledge or skill not related to the purpose of the test is required to answer an item correctly. 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 a cause of construct-irrelevant variance. Test takers whose metric conversion skills are weak may answer the item incorrectly, even though they could have successfully multiplied 1.8 times 1,000.

   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 variance depends on what is included in the intended construct.

[5] A source of construct-irrelevant variance can be a barrier to success. Cognitive barriers stem from differences in the knowledge bases of test takers. Affective barriers stem from differences in the emotional reactions of test takers. Physical barriers stem from differences in the abilities of test takers to see and hear tests and related material.

2. Avoid affective sources of construct-irrelevant variance.

   Affective sources of construct-irrelevant variance occur if language or images cause strong emotions that may interfere with the ability to respond to an item correctly. For example, offensive content may make it difficult for 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 variance. 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 controversial material.

   Even if test takers' performance is not directly affected, the inclusion of content that appears to be offensive, upsetting, controversial, or the like may lower test takers' and score users' confidence in the test and may lead people to believe that ETS tests are not fair.

3. Avoid physical sources of construct-irrelevant variance.

   Physical sources of construct-irrelevant variance occur (most often for test takers with disabilities) if aspects of tests interfere with the test takers' ability to attend to, see, hear, or otherwise sense the items or stimuli. For example, test takers who are visually impaired may have trouble understanding a diagram with labels in a small font, even if they have the KSAs that are supposed to be tested by the item based on the diagram.

Applications of the Guidelines

Application to ETS Tests: ETS requires a formal, documented fairness review of items and stimulus material for compliance with these guidelines. The reviews must be done by specially trained reviewers before the items are administered to 50 or

more test takers. With the exception of the section on additional requirements for K–12 tests, these guidelines apply to cognitive and noncognitive items and stimuli in all ETS tests, pretests, equating sets, pilot tests, field tests, and sample tests. Items in parent guides, student guides, test-preparation material, test bulletins, and items embedded in other communications also require fairness review. For fairness review of ETS publications other than tests, please see *ETS Guidelines for Fairness Review of Communications* (2009).

Application to Groups: While these guidelines apply to all test takers, the groups of primary concern are defined by the following.

- Age
- Disability
- Ethnicity
- Gender
- National or regional origin
- Native language
- Race
- Religion
- Sexual orientation
- Socioeconomic status

Other groups may be of primary concern in tests made specifically for use in countries other than the United States.

Application to Content and Skills Tests: 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 knowledge in a specific subject-matter area, such as biology, dance, English literature, nursing, or psychology.

Skills tests rarely need to include material on a specific topic for valid measurement. Therefore, skills tests are not likely to need any material for valid measurement that is out of compliance with these guidelines. Content tests, however, may have to include material on a specific topic that is important for valid measure-

ment of the tested subject, even if that material would otherwise be out of compliance with these guidelines.

Application to Tests for Different Countries or in Different Languages: Whether tests are in English or in some other language, these guidelines apply as written to tests designed primarily for use in the United States, though those tests may also be used worldwide.

Tests designed specifically for test takers in a particular country or area 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 tests are designed. For example, tests designed for use in Qatar and tests designed for use in Japan would require guidelines for what is considered offensive that are different from each other and different from the guidelines for tests designed for use in the United States. Similarly, tests designed for use in Europe and tests designed for use in Africa would have different guidelines for the representation of diverse people. Please see *ETS International Principles for Fairness Review of Assessments* (2009) for information on how to adapt the guidelines for use with tests made for particular countries other than the United States.

Application at Test Design: Concern with fairness in assessment begins as tests are being designed. When a construct can be measured in different ways that are reasonably equally valid and practical, consider these guidelines in determining how best to measure the construct.

GUIDELINE 1

AVOID COGNITIVE SOURCES OF CONSTRUCT-IRRELEVANT VARIANCE

Purpose: 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. Guideline 1 requires the avoidance of construct-irrelevant knowledge or skills.

The categories listed below are likely to require construct-irrelevant knowledge or skills that are unfair sources of difficulty for people in various groups. Avoid these content areas unless the knowledge required is relevant to the construct. Any of these content areas may be included if the knowledge is specifically related to the purpose of the test.

Unnecessarily Difficult Language: Avoid unnecessarily difficult language. Use the most accessible level of language that is consistent with valid measurement.[6] While the use of accessible language is particularly important for test takers who have limited English skills, the use of accessible language is beneficial for all test takers when linguistic competence is construct irrelevant. *ETS Guidelines for Using Accessible Language in Tests*, attached as appendix 1, provides additional information about how to avoid unnecessarily difficult language.

Difficult words and language structures may 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. Com-

[6] These guidelines do not require that tests be written in languages other than English.

plicated language structures may be appropriate if the purpose of the test is to measure the ability to read challenging material.

Avoid unnecessarily specialized vocabulary unless such vocabulary is important to the construct being assessed. What is considered unnecessarily 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. Please see Guideline 2 for more information about affective sources of construct-irrelevant variance.

Table 1 provides examples of common words that are generally acceptable and examples of specialized words that should be avoided unless they are important for validity. The words are within several content areas known to be likely sources of construct-irrelevant knowledge.

Topics Likely to Be Sources of Construct-Irrelevant Variance: Avoid requiring specialized knowledge that is unrelated to the purpose of the test to answer an item correctly. The following topics are likely sources of construct-irrelevant knowledge. Do not require specialized knowledge of these topics unless it is important for valid measurement. Aspects of the topics that are common knowledge expected of all intended test takers are acceptable, and mention of the topics in material that focuses on other topics is acceptable.

> **Military topics.** Unless it is important for validity, avoid requiring knowledge of military topics, such as conflicts, wars, battles, and military strategy. Avoid requiring knowledge about how military organizations function. Avoid requiring knowledge of weapons, of the functions of parts of weapons, or of how weapons work.
>
> A mention of war, weapons, or any other military topic in an item or stimulus primarily concerned with a different topic is acceptable. For example, a passage about Horatio Gates

Table 1. Examples of Common and Specialized Words

Field of Knowledge	Common Words	Specialized Words
Digital technology	Email, computer, Internet	JPEG, MP3, server[7]
Farming	Field, harvest, plow	Combine, thresher
Finance	Tax, salary, income	Arbitrage, hedge fund, preferred stock, tax-free bonds, venture capital
Law	Judge, jury	Subpoena, tort
Politics	President, vote	Earmark, filibuster, pork barrel
Science	Microscope, thermometer, degree	Byte, lumen, vacuole
Tools and machinery	Hammer, engine	Torque, flange, chuck
Transportation	Train, car, sail	Catamaran, drone, boom

that mentions he was a general in the American Revolutionary War is acceptable. It is acceptable to mention his role in the Battle of Saratoga. A passage that details the tactics that Gates used in the Battle of Saratoga, however, is not acceptable in a skills test. Military topics are acceptable when they are construct- relevant, as in some history tests.

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

[7] Terms in this field are likely to change status from specialized to common fairly rapidly.

“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.

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 the test taker as part of his or her education in art or literature, not as part of his or her education in religion. For more information about including religion in ETS tests, please see Guideline 2.

Specialized tools. Test items in skills tests should not require specialized knowledge; for example, knowledge of the purpose of a die, a compressor, or a router. Also, avoid requiring knowledge of how tools and machines work or are assembled or knowledge of the functions of various parts of tools or machines unless it is important for valid measurement.

Sports. If specialized knowledge of a sport is needed to answer an item, the item should not appear in a general skills test. For example, in a general skills test, do not require test takers to know how many points a field goal is worth in American football. Such an item would, however, be perfectly appropriate in a licensing test for physical education teachers in the United States.

United States culture. ETS tests are taken in many countries. Even tests limited to 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 of such topics as brands of products, customs, geography, government, history, holidays, institutions, laws, measurement systems (degrees Fahrenheit, inches, pounds, quarts, etc.), plants, politicians, political systems, public figures, slang, sports, television shows, or wildlife specific to the United States.

Do not assume that all test takers are from the United States. Do not use the word “America” to refer solely to the United States of America, and do not use the phrase “our country” to refer to the United States of America unless the context makes the meaning clear. Similarly, unless the context makes it clear, do not use the term “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. Do not assume that all test takers share the point of view that the United States has taken on international controversies.

GUIDELINE 2

AVOID AFFECTIVE SOURCES OF CONSTRUCT-IRRELEVANT VARIANCE

Purpose: The primary purpose of Guideline 2 is to avoid emotional reactions caused by inappropriate, construct-irrelevant test content that may affect people in different groups in different ways.[8] No group of test takers should have to face language and images that are unnecessarily contemptuous, derogatory, exclusionary, insulting, or the like. The contents of tests should not distract a test taker from the task of understanding a stimulus or responding to an item by unnecessarily inducing negative emotions. In addition, construct-irrelevant aspects of a test should not make test takers feel alienated or uncomfortable. It is important to avoid construct-irrelevant content that is commonly believed to be unfair, even if it has not been proven that test takers' performance is actually affected by such content.

Interpreting the Guideline: How controversial, inflammatory, offensive, or upsetting does material have to be in order to violate Guideline 2? Drawing a clear line is difficult because what is considered inappropriate will vary. Material that is acceptable to some groups may be offensive to other groups. Furthermore, the need for valid measurement and the way topics are treated must be taken into account when determining whether the material is acceptable. Therefore, judgment will always be required in interpreting Guideline 2. Consider the following factors in deciding whether material is in compliance with the guideline.

> **Age and sophistication of test takers.** Guideline 2 is to be interpreted most strictly for K–12 tests. Please see Additional Requirements for Fairness Review of K–12 Tests for the application of Guideline 2 to tests designed

[8] These guidelines do not apply to emotional reactions to the construct-relevant aspects of a test. For example, these guidelines are not violated if some test takers feel anxious attempting to solve valid problems in a mathematics test.

for school children in particular states, cities, or school districts. In general, the older and more sophisticated the test takers, the more liberally the guideline should be interpreted. For example, material that is unacceptable for high school students may be acceptable for college students.

Previous experience of test takers. Consider the kinds of material that test takers are likely to have been exposed to extensively in deciding whether some test material is likely to offend or upset them. Brief exposure does not justify the inclusion of upsetting material in a test. If, however, people have become accustomed to the material, it is not likely that encountering it again in a test would be excessively problematic.

Directness of the material. Almost any item that is set in some innocuous context could be construed as unfair for test takers who had undergone an unpleasant experience in a similar context, even though that experience is not directly addressed in the item. For example, a mathematics item about the average speed of a car could be construed as upsetting for test takers who have been involved in a car accident, even though the item does not specifically say anything about car accidents.

Contrast innocuous contexts with contexts that directly mention the potentially upsetting experience. For example, a mathematics item in a skills test about the average number of children killed per year in car accidents would be inappropriate for test takers, regardless of whether they have been involved in a car accident.

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. Directly problematic topics, such as those discussed below in Topics of Concern, are likely to be upsetting, even in the

absence of such scenarios. Those topics should be avoided unless they are important for valid measurement.

Importance for validity. Some reasonably controversial material may be important for validity, even in skills tests. For example, 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 Guideline 2 to illustrate certain derogatory attitudes commonly held in the past. A literature test for upper-level college students may appropriately include material that would be excluded from a test for younger test takers.

Sometimes the judgment about whether material is important for validity has to be made on the basis of an entire test rather than an individual item. For example, representation of the important aspects of the curriculum in a content test is a property of a whole test, not of a single item.

If controversial material is important for valid measurement, make clear to test takers that such material does not represent the views of ETS. This can be done in a variety of ways, such as by using quotation marks, indicating when a passage was originally written or when a cartoon was drawn, identifying the author or artist, or stating that the material does not represent the views of ETS or the testing program.

Topics of Concern: No topic is ever completely excluded from ETS tests because any topic that is important for validity may be tested. As specified below, however, some topics are best avoided unless they are important for valid measurement. Other topics need not be avoided completely but must be treated in as balanced, sensitive, and objective a manner as is consistent with valid measurement. Some stimuli and some items may neces-

sarily focus on problematic issues. Present such material in a way that will reduce its emotional impact.

Current events can cause new topics to be added to the list at any time. Therefore, any list of troublesome topics can only be illustrative rather than exhaustive. A topic is not necessarily acceptable merely because it has not been included on this list. It is a good practice to obtain a preliminary fairness review of potentially problematic material before time is expended on developing it.

Accidents, illnesses, or natural disasters. Avoid 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. Do not use test content to advocate any particular cause or ideology. Items and stimulus material should be neutral and balanced whenever possible. Do not 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.

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, so it is preferable not to focus on any aspects of the topic unnecessarily. 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.

Evolution, however, is a core concept in biological science. Furthermore, topics associated with evolution are important in several other disciplines. Therefore, 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. Some states even restrict topics associated with evolution (such as fossils or the age of Earth). Please see Additional Requirements for Fairness Review of K–12 Tests for more information.

Group differences. Avoid 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 caring for others, courage, honesty, trustworthiness, physical attractiveness, quality of culture, and so forth. Do not treat any one group as the standard of correctness against which all other groups are measured.[9] 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 who do not understand the joke may be offended. Similarly, treat irony and satire very carefully because some people could take such material literally and be offended by it. In particular, avoid construct-irrelevant humor that is based on disparaging any groups of people, their strongly held beliefs, their concerns, or

[9] This is not intended to prohibit the use of reference groups in statistical analyses.

their weaknesses. It is acceptable to test understanding of humor, irony, and satire when it is important for valid measurement, as in some literature tests.

Images for international populations. 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 the following types of images or descriptions unless important for valid measurement.

- People dressed in tight or revealing clothing
- People who are posed immodestly
- Men and women touching each other
- Men and women together in intimate settings such as a dormitory room

Certain hand signals that are acceptable in the United States are inappropriate in some other countries. Avoid images of 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 curled), 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 cultural expectations. People intended to be professors, for example, should look older than the students depicted and be dressed conservatively. People intended to be students should not be shown in excessively casual dress or behaving disrespectfully in the presence of an authority figure.

Inadvertent 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 avoid them in tests unless they are construct relevant. Cryptic references can be a problem because test developers may be unaware of the references. For example, the time

4:20 has become associated with drug use, the date April 20 (Hitler's birthday) has become associated with White supremacists, and the number 311 (three times K, the eleventh letter of the alphabet) has become associated with the Ku Klux Klan.[10]

Avoid pictures and diagrams that include inappropriate symbols, such as swastikas, unless they are important for valid measurement. Avoid using the names of people associated with inappropriate or criminal behavior unless they are important for valid measurement. Avoid construct-irrelevant double entendres.

Luxuries. Avoid depicting situations that are associated with spending money on luxuries, such as eating in exclusive restaurants, joining a country club, taking a cruise, buying a swimming pool, owning an expensive car, having a private trainer, going on a ski trip, and the like, unless doing so is important for validity.

Personal questions. Unless important for validity, avoid asking test takers to respond to excessively personal questions regarding themselves, family members, or friends.[11] Topics such as the following are generally considered inappropriate.

- Antisocial, criminal, or demeaning behavior
- Family or personal wealth[12]
- Political party membership
- Psychological problems
- Religious beliefs or practices or membership in religious organizations
- Sexual practices or fantasies

[10] The Fairness Steering Committee will share information about inappropriate, cryptic references to help test developers identify them. Please send examples of such references to the Fairness Steering Committee.

[11] 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.

[12] Questions about wealth may be acceptable if the information is required to determine qualification for some program or benefit.

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.

Passing 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 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. Do not use phrases closely associated with religion as figures of speech. For example, do not use the phrase “born again” as a general intensifier or use “cross to bear” to stand for a person’s problem. Do not use “crusade” or “crusader” outside of their historical context. 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 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.

Slavery. Except when important for valid measurement in content tests, slavery should not be the main focus of any material. Mention of the topic in skills tests is acceptable if the emphasis of the material is on something else. For example, a passage that focused on the accomplishments of Frederick Douglass or a passage that focused 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 might be acceptable in a history test.

Though “slave” is still an acceptable term, “enslaved person” is preferred by some. Use “slaveholder” rather than “slave owner.” When writing about the era in which slavery ended in the United States, refer to “former slaves” and “freed people.” Do not refer to “freed slave” or “free Black” unless quoting historical or literary material. Use “freedmen” only when quoting historical or literary material or with respect to the names of organizations.

Societal roles. Take care to demonstrate that people in different groups are found in a wide range of societal roles and contexts. Avoid language and images that suggest that all members of any single group are people in higher-status positions or lower-status positions. For example, do not have all the Black people represented in a test be of lower status than the White people represented in the test. Do not have all the executives represented in a test be male and all the support staff be female. Do not overrepresent members of any group in

examples of inappropriate, foolish, unethical, or criminal behavior.[13]

Stereotypes. Avoid stereotypes (both negative and positive) in language and images unless important for valid measurement. Do not attribute characteristics to individuals on the basis of group membership (unless the group was composed on the basis of that characteristic). Avoid using phrases that encapsulate stereotypes, such as "Dutch uncle," "Indian giver," "women's work," or "man-sized job." Avoid use of words such as "surprisingly" or "amazingly" when the surprise or amazement is caused by a person's behavior contrary to a stereotype. For example, avoid such sentences as "Surprisingly, a girl won first prize in the science fair."

Avoid passages about stereotypes in skills tests. Avoid stereotypes in tests as sources of wrong answer choices. Test takers who select a wrong 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 group members 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.

[13] Some cultures are not egalitarian and would not, for example, have women in higher-status positions. 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).

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 to glorify suicidal behavior or other self-destructive behavior.

Topics best avoided. Some topics, such as the following and others similarly problematic, are best avoided unless they are important for valid measurement in content tests.

- Abortion
- Abuse of people (especially children) or animals
- Atrocities or genocide
- Contraception
- Euthanasia
- Experimentation on human beings or animals that is painful or harmful
- Hunting or trapping for sport
- Rape
- Satanism
- Torture
- Witchcraft

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 sixty-five 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 the detailed effects of violence, or on suffering unless important for valid measurement. Violence and suffering are too pervasive 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 animals are depicted eating other animals. Do not, however, dwell unnecessarily on the gruesome or shocking aspects of violence and suffering.

Use Appropriate Terminology for Groups: If group identification is necessary, it is generally most appropriate to use the terminology that group members prefer. Do not use derogatory names for groups in ETS tests, even if some group members use them to refer to other members of the group.

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 or areas other than the United States, determine the client’s preferences concerning terminology. It is not, however, appropri-

ate to use derogatory terms, even if a client should state such a preference. It is acceptable for material that is clearly quoted from external sources to use terms other than those listed here, but inflammatory terms should be avoided.

African American people. 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, avoid the use of “black” as a negative adjective, as in “black magic,” “black day,” or “black hearted.”

Asian American people. The terms “Asian American,” “Pacific Island American,” and “Asian/Pacific Island American” should be used as appropriate. The term “Asian” includes people from India. If possible, use specific terminology such as “Chinese American” or “Japanese American.” Do not use the word “Oriental” to describe people unless quoting historical or literary material or using the name of an organization.

People who are bisexual, gay, lesbian, or transgendered. Identify people by sexual orientation only when it is relevant to the construct to do so. Do not use the labels gratuitously.

The words “bisexual,” “gay,” “lesbian,” and “transgendered” are all acceptable. 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. Use the phrase “sexual orientation” rather than “sexual preference.”

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 preferred usage is to put the person first and the disabling condition after the noun, as in “a person who is blind.”

Avoid terms that have negative connotations or that reinforce negative judgments (e.g., “afflicted,” “crippled,” “confined,” “inflicted,” “pitiful,” “victim,” or “unfortunate”). 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.” Avoid euphemistic or patronizing terms such as “special” or “physically challenged.” Avoid the use of such words and phrases as “inspirational,” “courageous,” achieving success “in spite of” a disability, and “overcoming” a disability.

Do not use 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. Avoid references such as “invalid,” “sickly,” or “victim.” 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.

Table 2 distinguishes among appropriate and inappropriate terms related to disabilities.

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.

Hispanic American people. The terms “Latino American” (for men), “Latina American” (for women), and “Hispanic American” are acceptable and may be

Table 2. Use of Terms Related to Disabilities	
Abnormal, normal	Unacceptable in most cases. The terms “normal” and “abnormal” are not appropriate for referring to people except in biological or medical contexts.
ADA	Acceptable if one is referring to the Americans with Disabilities Act itself as a piece of legislation. Not acceptable when referring to people or tests. A person is not an “ADA candidate” but rather an “individual with a disability.”
Blind and visually impaired	The noun form, “the blind,” is not acceptable except in the names of organizations or in literary or historical material. The adjectival form, “a blind person,” is acceptable in subsequent references if “person who is blind” is used in the initial reference. Similarly, it is preferable to use “person who is visually impaired” rather than “visually impaired person” to put the emphasis on the individual rather than on the disability.
Deaf	Acceptable as an adjective, but sometimes the term “deaf” or “hard of hearing” may be used as a noun (e.g., School for the Deaf). 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” and “deaf mute.”
Down syndrome	Use the term “Down syndrome” rather than “Down's syndrome.” Avoid the obsolete and inappropriate term “Mongoloid.”
Hearing impaired	Generally to be avoided. The Deaf community and educators of individuals with hearing loss prefer “deaf and hard of hearing” to cover all gradations of hearing loss.

Interpreting	Acceptable. It describes a person (an interpreter) translating a signed language into a spoken language or vice versa.
Learning disabled	Acceptable as an adjective, but preferred usage is “a person with a learning disability.”
Mentally retarded	Acceptable, but preferred usage is “developmentally disabled” or “developmentally delayed.” Do not use the term “retarded” by itself. Outside of a technical context, do not use words such as “moron” or “idiot.”
Paraplegic, quadriplegic	Acceptable as adjectives, not as nouns.
Physical disability	Acceptable.
Spastic	Unacceptable to describe a person. Muscles are spastic, not people.

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.

Members of minority groups. Members of 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. Avoid the terms “disadvantaged minorities” and “underserved minorities.” The terms are vague and loaded with unstated assumptions. Use the names of the specific groups the terms are intended to include.

Multiracial people. The terms “biracial” and “multiracial,” as appropriate, are acceptable for people who identify themselves as belonging to more than one race. “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 such as the National Association for the Advancement of Colored People.

Native American people. The terms “American Indian” and “Native American” are acceptable. Avoid use of 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 “brave” (to refer to a Native American man) are not acceptable except in historical or literary material.

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, but it lacks specificity, as it may include all levels of competence ranging from people who have very limited English skills to people who are not distinguishable from native speakers. “English language learner (ELL)” is the preferred term for K–12 students who are not yet fully competent in English. The term “English as a second language (ESL)” applies to learning English in an English-speaking environment, whereas “English as a foreign language (EFL)” applies to learning English in a non-English-speaking environment. “Limited English proficient (LEP)” is generally used in a legislative context.

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 as an adjective to put the emphasis on the person rather than the person’s lack of English proficiency. However, once ELL has been used as an adjective, it is acceptable to refer to people as English language learners or ELLs.

Older people. It is best to refer to older people by specific ages or age ranges; for example, “people age sixty-five 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 below the poverty line. The term “poverty” is defined by the government of the United States. Use the term in accordance with that definition in tests designed primarily for use in the United States. Do not use the term “poverty-stricken.” Terms such as “low income” and “economically disadvantaged” are vague. The term “poor” is vague and ambiguous. It is preferable to use specific income ranges. Do not use “poor” as a noun except when quoting literary or historical material. Do not refer to people with the term “lower class” unless it is clearly defined or in literary or historical contexts. It is preferable to refer to levels of socioeconomic status than to levels of class.

White American people. 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 with “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.

Women and men. Women and men must 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 called 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. The terms “male” and “female” are acceptable but tend to be limited to scientific contexts and questionnaires.

Women must not be described by physical attributes when men are described by mental attributes or professional position; the opposite is also to be avoided. Gratuitous references to a person's appearance or attractiveness are not acceptable.

Women eighteen or older should be referred to as women, not girls. Men eighteen or older should be referred to as men, not boys.

Language that assumes that all members of a profession are either all males or all females is unacceptable. Generic terms such as “poet,” “doctor,” and “nurse” include both men and women; and modified titles such as “poetess,” “woman doctor,” or “male nurse” are not acceptable.[14] Role labels such as “scientist” or “executive” include both men and women. Do not use expressions such as “the soldiers and their wives” that assume

[14] “Actress” is acceptable when paired with “actor,” as in “awards for the best actor and actress.”

only men fill certain roles unless such is the case in some particular instance.

Do not couple generic role words with gender-specific pronouns or actions 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. Do not refer to objects using gender-specific pronouns except in historical or literary material.

Using “he” or “man” to refer to all people is not acceptable. Avoid the use of generic “he” or “man” unless it is included in historical or literary material. Alternating generic “he” and generic “she” is unacceptable because neither word should be used to refer to all people.[15] Avoid the constructions “he/she” and “(s)he.”

Table 3 gives examples of unacceptable usages of “man” and “he,” with acceptable alternatives.

Represent Diverse People: If a test mentions or shows people, test takers should not feel alienated from the test because no member of their group is included. Therefore, the ideal test would reflect the diversity of the test-taking population. While it is not feasible to include members of every group in a test, strive to obtain at least the level of diversity described below in tests that mention or show people.

Application of the Guideline: 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 or area other than the United States should be that of the test-taking population in the country or area for which the test is designed. Consult the client to determine the characteristics of the test-taking population to be reflected in a test made for a particular country or area. Also, consult the client to determine the

[15] When referring to a particular person as an example of people in a group, it is acceptable to refer to that person using the pronouns appropriate to his or her gender.

Table 3. Alternatives to Use of Generic Male References	
Unacceptable	**Acceptable**
Chairman	Chair, presiding officer, leader, moderator The terms “chairman” and “chairwoman” are acceptable when referring to specific men and women. Do not use “chairman” for a man along with “chairperson” for a woman; the terms are not parallel.
Fireman, mail-man, salesman, insurance man, foreman	Firefighter, mail carrier, sales representative, insurance agent, supervisor
If a student studies, he will learn. (Unacceptable when “a student” refers to all students)	If a student studies, she or he will learn. If a student studies, he or she will learn. If students study, they will learn. A student who studies will learn. Students who study will learn.
Man (as a verb)	Staff, provide workers
Mankind, man	Humanity, human beings, people
Manmade	Synthetic, artificial
Manpower	Workers, personnel, labor, workforce

characteristics of the test-taking population to be reflected in a test made specifically for a particular jurisdiction within the United States such as a state, city, or school district.

Gender balance. In skills tests, women and men should be reasonably equally represented. 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 content area. In occupational tests, it is appropriate to depart from the gender distribution of the members of the occupation to improve the gender balance of the test, as long as the departure is not so great as to disconcert the test takers.[16] Some of the men and women shown should be equivalent in status. For example, do not show all the doctors as male and all the nurses as female, or vice versa.

Racial and ethnic balance. For skills tests, strive to have about 20 percent of the items that mention people represent people from what are commonly considered to be minority groups in the United States or people from the countries of origin of those groups. For example, include African American people or African people, Asian American people or Asian people (including people from India), and Latino people from the United States or from Latin America. Also include indigenous groups from the United States or from other countries, such as the Guarani Indians of Paraguay. 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 the item by using names that are inordinately difficult for test takers to decode.

In some content tests, such as history tests or literature tests, the proportion of items dealing with diverse groups may be fixed by the test specifications. If the proportions

[16] Some representation of both genders is required, however.

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

For all tests, if it is compatible with valid measurement, try to achieve at least a rough parity in status of the people depicted in different racial and ethnic groups. Do not, for example, depict all managers as White and all workers as Black or Hispanic.

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 a stereotype concerning the lack of independence of people with disabilities. Ideally, the focus will be on the person, and the disability will be incidental rather than the focus of the image or text.

GUIDELINE 3

AVOID PHYSICAL SOURCES OF CONSTRUCT-IRRELEVANT VARIANCE

Purpose: The purpose of Guideline 3 is to help ensure that there are no unnecessary physical barriers in items or stimulus material (such as needlessly cluttered graphs) that may cause construct-irrelevant score variance, particularly for people with disabilities.[17]

This guideline applies to all tests and is not limited to the modified or alternative forms developed for people with certain disabilities.

Types of Barriers

Essential aspects. Some physical aspects of items are essential to measure the intended construct, even if they cause difficulty for 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. 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 items are helpful for measuring the intended construct, even if they cause difficulty for people with disabilities. For example, cartoons are often used as stimuli to elicit writing or speech in tests of English as a second language, even though the cartoons are physical barriers for test takers

[17] See the *ETS Guidelines for Test Accessibility* (2009) for additional information about making tests accessible for people with disabilities.

who are blind. Stimuli other than cartoons could be used in this case, so the cartoons are not essential. The cartoons, however, are very helpful as stimuli when it cannot be assumed that the test takers share a common native language. Fairness concerns about the helpful aspects of items should be raised at the design stage of test development when the item type is first considered. If decisions about the helpful aspects of items have been made and reviewed at the design stage and apply to an entire class of items, it is not appropriate to raise fairness challenges about those helpful aspects in later reviews of individual items.

Unnecessary aspects. Some physical barriers, however, 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. The focus of Guideline 3 is on the avoidance of unnecessary physical barriers in items and stimuli.

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.[18]

- Construct-irrelevant charts, maps, graphs, and other visual stimuli

[18] Material on Web sites must comply with legal requirements for accessibility. Consult with the office of the ETS General Counsel for the latest requirements.

- 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 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
- 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 labels used for options in the multiple-choice items based on the stimulus
- Letters that look alike (e.g., O, Q) used as labels for different things in the same item/stimulus
- Letters that sound alike (e.g., s, x) used as labels for different things in the same item/stimulus[19]
- 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)

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

- Special symbols or non-English alphabets (unless that is standard notation in the tested subject, such as Σ in statistical notation)
- Uppercase and lowercase versions of the same letter used to identify different things in the same item or stimulus (unless that is standard notation in the tested subject, 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 or helpful for valid measurement. Avoid them if they are unnecessary or not helpful.

In addition, ensure that recorded material is clear enough to avoid having the quality of the recording serve as a source of construct-irrelevant variance. 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 variance.

Reduce the need to scroll down the computer screen to access parts of stimulus material to the extent possible, unless the ability to scroll is construct relevant.

ADDITIONAL REQUIREMENTS FOR FAIRNESS REVIEW OF ETS K–12 TESTS

Rationale: The fairness review guidelines imposed by states, cities, or school districts for their K–12 tests are often more extensive than the requirements for other tests. [20] Jurisdictions are extremely cautious about the content of K–12 tests because 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 review guidelines.

The following requirements for K–12 tests are in addition to the fairness review guidelines that are to be followed for all ETS tests. 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 states.

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 three fairness review guidelines that apply to all ETS tests.

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 not appropriate for tests, even though they may be discussed in classrooms**.**

---

[20] K–12 tests are made for a particular state, city, or school district. 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. NAEP follows the additional guidelines for K–12 tests.

Additional Requirements: The following requirements apply to stimuli, items, and all art and graphics. The requirements are all extensions of Guideline 2 (avoid affective sources of construct-irrelevant variance). Any redundancy with the discussion of Guideline 2 is intended to stress the additional care to be employed in tests for young students. **Content that is required by a jurisdiction's content standards may be tested, even if it would otherwise be out of compliance with the additional requirements.**

Emotionally charged topics. Avoid concepts that may be emotionally charged for K–12 students. These include serious illnesses (e.g., cancer, HIV, AIDS, herpes, tuberculosis, smallpox, anthrax) and people or animals who are killed or are dying.

Avoid items or stimuli about natural disasters, such as earthquakes, hurricanes, tornadoes, floods, or forest fires, unless the disasters are treated as scientific subjects and there is little or no mention of the destruction caused and loss of life.

Avoid mention of divorce, loss of jobs, layoffs, and other family situations that students may find upsetting.

Avoid depictions or suggestions of interpersonal violence or disharmony, including playground arguments, fights among students, bullying, cliques, and social ostracism. Avoid dissension, or even suggestion of dissension, among family members or between students and teachers.

Avoid depictions of graphic violence in the animal kingdom. Avoid a focus on pests (e.g., rats, roaches, and lice) or on creatures that may be frightening to students (e.g., scorpions, poisonous snakes, and spiders). Any animals that are important for valid measurement may be used, however.

Certain animals may be sensitive topics for specific cultural groups (e.g., the owl for some Native American

nations). Check these issues with the client or the client's bias committee.

Offensive topics. Avoid topics that may be offensive to particular groups in a jurisdiction. Such topics include drinking alcohol, smoking or chewing tobacco, and gambling. Avoid suggestions of drug use, including the use of prescription drugs. Other topics that may be inappropriate for specific groups of students include birthday celebrations, Christmas, Halloween, and Valentine's Day. The subject of dancing, including school dances (such as proms), may be offensive to some groups. Some clients may prefer that mention of particular kinds of music (e.g., rap, rock and roll) be avoided. Some clients may request that going to the movies not be mentioned.

Avoid references to a deity, including expressions like "Thank God." Also avoid euphemisms for such references (for example, "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.

Do not discuss or refer to extrasensory perception, UFOs, the occult, or the supernatural.

Avoid texts that are preachy or moralistic, as they may offend populations that do not hold the values espoused.

Controversial topics. There are many topics that are controversial and thus should be excluded from K–12 testing in addition to those discussed in Guideline 2. Such topics include gun control, welfare, global warming, the suffering of individuals at the hands of a prejudiced or racist society, a focus on individuals overcoming prejudice, and the specific results of discrimination against women.

The topic of evolution, with associated topics of natural selection, fossils, geologic ages (e.g., millions of years

ago), dinosaurs, and similarities between people and primates, should not appear in K–12 testing unless required by specific content standards.

Do not appear to promote or defend particular personal or political values in discussions of, for example, protection of the environment, deforestation, or labor unions. Maintain a neutral stance on controversial issues, even when content standards include those issues. A possible exception could be standards requiring passages or stimuli that are designed to be persuasive. Such passages or stimuli should be clearly labeled as persuasive or editorial text.

Avoid biographical passages that focus on individuals who are readily associated with offensive topics. It is best to avoid biographical passages that focus on individuals who are still living. Their future actions or activities are unpredictable and may result in fairness problems.

Inappropriate behaviors. Do not use material that models or reinforces inappropriate student behaviors. Students should not be shown playing tricks on or trying to deceive teachers or other adults. Items or stimuli should not show anyone lying, stealing, or running away from home, or even considering those behaviors.

Avoid topics such as going without sleep, failing to attend school or do homework, and eating large quantities of junk foods.

Do not appear to recommend that students violate good safety practices (e.g., keeping dangerous animals, entering homes of unknown adults), even if everything turns out well. Avoid suggestions of sexual activity. Avoid words or phrases that carry a sexual connotation.

Do not portray children coping with adult decisions or situations (e.g., caring for siblings, supporting the emotional needs of a parent).

Do not express or imply cynicism about charity, honesty, or similar values.

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 to avoid. 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 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 specifically required by a content standard. The subject of dancing, while otherwise to be avoided, may be appropriate in fine arts assessments.

If such topics as war, slavery, or disease are required by the content standards, the topics should be presented in a manner sensitive to the feelings of individual students who may have strong emotions concerning those issues.

CONCLUSION

Validity is linked to the quality of the inferences made about test takers on the basis of their scores. The purpose of fairness review is to identify sources of construct-irrelevant variance that might plausibly impair the quality of the inferences made about members of certain groups of test takers. These guidelines are meant to help the people who design, develop, and review ETS items and tests make fair and valid tests.

It is impossible, however, to develop rules and examples for fairness review that will cover every situation. Furthermore, what is considered fair changes over time, so some aspects of these guidelines will eventually become obsolete, and other guidelines may be added. In some cases, application of the guidelines will require the careful evaluation of competing priorities. Both excessively zealous and excessively lax interpretations of these guidelines are counterproductive.

In 1999 the *Standards for Educational and Psychological Testing* (AERA, APA, NCME) stated, “It is unlikely that consensus in society at large or within the measurement community is imminent on all matters of fairness in the use of tests.” Consensus on all matters of fairness is still unlikely, and professional judgment will always be required in the course of fairness review. These guidelines should help authors, editors, and fairness reviewers apply their judgment to make ETS tests as valid and as fair as possible for all of the people who take them.

APPENDIX 1

ETS GUIDELINES FOR USING ACCESSIBLE LANGUAGE IN TESTS

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 groups of test takers may in particular 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.[1] It is a practice that has the potential to minimize construct-irrelevant variance for all test takers.

On some programs, clients may have specific style guides or other policies related to language use. These guidelines are not intended to conflict with such client preferences. However, ETS staff should seek to maximize the accessibility of language used in all programs. This effort may include advising clients on the value of making language accessible in order to minimize construct-irrelevant variance.

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.).

The guidelines permit language that is part of the construct being tested to 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.

---

[1] For more information on assessing English language learners and students with disabilities, see *ETS Guidelines for the Assessment of English Language Learners* (2009) and *ETS Guidelines for Test Accessibility* (2009), respectively.

The following are some specific examples of the exception for test content.

In a reading-comprehension test, the reading passages should be as challenging as is required for valid measurement. For example, a reading test for admission to college must include college-level passages.

- In a test of vocabulary or language structure, the words or language structures should be as challenging as needed to meet the test specifications.

- Subject-matter tests may use difficult vocabulary and language structures that are part of the subject matter. Indeed, mastery of the vocabulary of a field can be an essential part of understanding that field. It is entirely appropriate, therefore, to assess knowledge of words such as “ontogeny” on a biology test or words such as “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 non-construct-related aspects of the test material should use the most accessible level of language that is consistent with validity.[2]

Guidelines for Accessible Language: Writing in a clear and accessible way is a complex and often subtle craft. In striving to cast tests in the most accessible language possible, consider the audience and the construct being assessed and continually look

[2] These guidelines do not require that tests be written in languages other than English.

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 meaning. 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 complex and difficult to follow.

- Make the referent of a pronoun as clear as possible. Usually, the referenced noun should be the closest one 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”).[3]

- 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 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 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

[3] Particularly at the K–12 level, vocabulary guidelines such as Mogilner et al., *Children’s Writers’ Word Book* (2006), and EDL, *EDL Core Vocabularies* (1989), can be useful resources.

words with multiple meanings is important to the construct).

- When using words in a part of speech that is not common for the word (e.g., “foot” as a verb), take care to ensure that the context makes the intended meaning clear.

- Use personal pronouns when they help with communication. When appropriate, in directions address the reader as “you” rather than using a more abstract, impersonal reference such as “one.” Avoid using “you” in test questions that are supposed to have a single best key, however.

If a passage contains challenging vocabulary that is not part of the construct and that cannot be edited, consider glossing or footnoting the difficult vocabulary. 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, “Toni Morrison wrote *The Bluest Eye*” is preferred to “*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 on some points that are specific to the genre of the test item.

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. 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 he 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. He sold three of the sheep. Which number sentence represents how many sheep he 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. If the construct being assessed involves extracting relevant information from irrelevant information, clear guidelines for the inclusion of irrelevant information should be established.

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 will often be accessible to a wider range of students than a home-based context.

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.

Critical Reading—College Placement

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 first sentence, the word “panacea” means
Comment: The less accessible version introduces extraneous language. The more accessible version is succinct.	

Analytic Writing—Graduate Admissions

A writing prompt containing the expression “moderation in all things” resulted in a number of ELL test takers who were unfa-

miliar with the term “moderation” basing their essays on similar words (“modesty,” “modification,” etc.). This type of problem can be avoided either by avoiding key words that are likely to be unfamiliar to a number of test takers or by defining or glossing such words to help ensure comprehension.

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.	

Less Accessible	More Accessible
When Ms. Johnson pulled her car into the parking garage, she received a ticket stamped with the time 11:12 A.M. When she left the garage that afternoon, the time was 2:15 P.M. What was the total length of time that Ms. Johnson’s car was in the parking garage?	Amy 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 (a parking garage in which times are stamped on a ticket) and 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

ADDITIONAL FAIRNESS ACTIONS

Fairness review is only one of the ways in which ETS strives to make tests fair for all intended test takers. ETS strives to ensure fairness throughout the life of a test by requiring the following.

Impartiality: An important aspect of fairness is treating people impartially, regardless of characteristics that are not relevant to the test being given (such as gender, race, ethnicity, or disability). ETS gives all test takers respectful treatment, equal access to relevant testing services, and useful information about the assessment. ETS maintains standardization of registration, administration, and scoring to ensure that all people are treated appropriately. As part of treating test takers appropriately, ETS provides reasonable accommodations for people with disabilities to help ensure that the test is measuring relevant knowledge and skills rather than the effects of a person's disability.

External Contributors: ETS requires 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 men and women 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 the way people in different groups perform on each test item are used. The statistics measure differential item functioning, or DIF. 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. The statistics are applied whenever sample sizes are large enough to permit their use. 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. For more information about DIF, see Dorans and Holland (1993) and Zieky (1993).

Validation: A crucial aspect of fairness is validation. Essentially, validation is the collection of evidence to determine whether the inferences made on the basis of test scores are appropriate. Careful validation is done for every ETS test. 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. For more information about validity, see Messick (1989).

Test Interpretation and Use: Even a fair test can be used unfairly. For example, interpreting scores as measures of innate ability, when the opportunity to learn the tested material is not equally distributed, is unfair. ETS specifies the appropriate interpretation and use of its tests and makes the information available to score recipients.

Research: Finally, ETS supports a great deal of research directly related to test fairness and has been doing so for many years. For example, ETS researchers in the 1960s used statistical techniques to investigate the relationship between performance on test items and group membership (Cardall and Coffman, 1964). DIF statistics were developed at ETS in the 1980s (Dorans and Holland, 1993). Messick (1989) clarified the strong links between validity and fairness, and Willingham and Cole (1997) made a major contribution to the understanding of gender differences in assessment results. Barton (2003) and Barton and

Cooley (2008) explored the reasons for the differences in test scores between Black and White test takers. Dorans and Liu (2009) researched the extent to which test scores are related to each other in the same way across different groups of test takers. Pitoniak and others (2009) prepared research-based guidelines for the appropriate assessment of English language learners. For reports of ETS research related to fairness, please visit www.ets.org. The ongoing research helps ETS to ensure that its tests are as fair as possible.

SOME USEFUL REFERENCES ON FAIRNESS

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). 1999. *Standards for educational and psychological testing.* Washington, DC: AERA, APA, and NCME.

American Psychological Association. 1977. *Guidelines for non-sexist language in APA journals.* Washington, DC: APA.

American Psychological Association. 2001. *Publication manual of the American Psychological Association.* Washington, DC: APA.

Angoff, W. H. 1993. Perspectives on differential item functioning methodology. In *Differential item functioning*, ed. P. Holland and H. Wainer. Hillsdale, NJ: Lawrence Erlbaum.

Associated Press. 2007. *The Associated Press stylebook*. New York: Basic Books.

Barton, P. 2003. *Parsing the achievement gap*. Princeton, NJ: ETS.

Barton, P., and R. Coley. 2007. *The Family: America's smallest school*. Princeton, NJ: ETS.

Barton, P., and R. Coley. 2008. *Windows on achievement and inequality*. Princeton, NJ: ETS.

Berk, R. A., ed. 1982. *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

Camilli, G. 2006. Test fairness. In *Educational measurement*, ed. R. Brennan. Westport, CT: American Council on Education / Praeger.

Cardall, C., and W. Coffman. 1964. *A method for comparing the performance of different groups on the same items of a test* (RR 9, 64–65). Princeton, NJ: ETS.

Cole, N. S., and P. A. Moss. 1989. Bias in test use. In *Educational measurement*, ed. R. L. Linn. Washington, DC: American Council on Education.

Cole, N. S., and M. J. Zieky. 2001. The new faces of fairness. *Journal of Educational Measurement* 38: 4.

Dorans, N. J. and P. W. Holland.1993. DIF detection and description: Mantel-Haenszel and standardization. In *Differential item functioning*, ed. P. Holland and H. Wainer. Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., and J. Liu. 2009. *Score equity assessment: Development of a prototype analysis using SAT Mathematics test data across several administrations* (RR 8, 2009). Princeton, NJ: ETS.

Educational Testing Service. 2002. *ETS standards for quality and fairness*. Princeton, NJ: ETS.

Educational Testing Service. 2003. *ETS fairness review guidelines*. Princeton, NJ: ETS.

Educational Testing Service. 2007. *ETS international principles for fairness review of assessments*. Princeton, NJ: ETS.

Educational Testing Service. 2009. *ETS guidelines for fairness review of communications*. Princeton, NJ: ETS.

Educational Testing Service. 2009. *ETS guidelines for test accessibility*. Princeton, NJ: ETS.

Holland, P. W., and N. J. Dorans. 2006. Linking and equating. In *Educational measurement*, ed. R. L. Brennan. Westport, CT: American Council on Education / Praeger.

McGraw-Hill. 1983. *Guidelines for bias-free publishing.* New York: McGraw-Hill.

Messick, S. 1989. Validity. In *Educational measurement*, ed. R. L. Linn. Washington, DC: American Council on Education.

Petersen, N. S., and M. R. Novick. 1976. An evaluation of some models of culture-fair selection. *Journal of Educational Measurement* 13, 3.

Pitoniak, M., J. Young, M. Martiniello, T. King, A. Buteux, and M. Ginsburgh. 2009. *Guidelines for the assessment of English language learners.* Princeton, NJ: ETS.

Ramsey, P. 1993. Sensitivity review: The ETS experience as a case study. In *Differential item functioning*, ed. P. Holland and H. Wainer. Hillsdale, NJ: Lawrence Erlbaum.

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

Thompson, S. J., C. J. Johnstone, and M. L. Thurlow. 2002. *Universal design applied to large scale assessments.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Willingham, W. W., and N. S. Cole. 1997. *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Zieky, M. 2006. Fairness reviews in assessment. In *Handbook of test development*, ed. S. Downing and T. Haladyna. Mahwah, NJ: Lawrence Erlbaum.