

Educational Testing Service
FAIRNESS REVIEW GUIDELINES

Copyright © 2003 by Educational Testing Service. All rights reserved.

Contents

PREFACE	1
INTRODUCTION	4
GUIDELINES	8
Guideline 1. Treat people with respect in ETS materials.....	8
Guideline 2. Minimize the effects of construct-irrelevant knowledge or skills.	9
Guideline 3. Avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting.....	12
Guideline 4. Use appropriate terminology.	15
Guideline 5. Avoid stereotypes.	20
Guideline 6. Represent diversity in depictions of people.	22
GENERAL PROCEDURES	24
Essential Aspects of Fairness Review.....	24
Preliminary Review.....	25
Resolution of Disputes.....	25
Need for Additional Fairness Review.....	26
SOME USEFUL REFERENCES ON FAIRNESS IN ASSESSMENT.....	27

PREFACE

I am proud to issue the updated and enhanced version of the *Educational Testing Service Fairness Review Guidelines*. The *Guidelines* are integral to the achievement of that part of ETS's corporate mission committing ETS "to help advance quality and equity in education by providing fair and valid assessments."

Fairness review. Fairness review is intended to identify and remove invalid aspects of test questions that might hinder people in various groups from performing at levels that allow appropriate inferences about their relevant knowledge and skills.

Fairness review in context. At ETS, mandated reviews for the fairness of test questions and other materials are conducted by specially trained reviewers who follow the guidelines and procedures described in this document.

Fairness review, however, is only one of the ways in which ETS staff strive to make their tests fair for everyone. ETS requires attention to fairness throughout the life cycle of a test. Testing programs must demonstrate that reasonably anticipated potential areas of unfairness will be addressed as the test is designed, developed, administered, and scored, and as the results are used.

In addition to fairness review, ETS requires equality of treatment for all test takers, diverse external contributors to test development, the use of differential item¹ functioning statistics, validation of tests, and the provision of information about appropriate test interpretation and use.

- **Equality of treatment.** An important aspect of fairness is treating people with impartiality regardless of such characteristics as gender, race, ethnicity, or disability that are not relevant to the test being given. All test takers are given 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 equally. People with disabilities are provided accommodations to help ensure that the test is measuring the relevant knowledge and skill rather than the effects of the 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 test development committees to determine the knowledge, skills, and abilities to be tested. Committee members may also review, revise, and select the questions 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 question writers and reviewers, and as test reviewers.

¹ An "item" is a test question.

- Differential item functioning. An empirical measure of the way matched people in different groups perform on each test question, called differential item functioning or DIF, is used whenever sample sizes permit as an empirical check on the fairness of questions. DIF occurs when people in different groups perform in substantially different ways on a test question, even though they have very similar scores on the test. 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. Questions with high DIF are reviewed for fairness and removed before the test is scored, if the questions are judged to be unfair.
- Validation. A crucial aspect of fairness is validation. Essentially, validation is the collection of evidence to determine whether or not the inferences made on the basis of test scores are appropriate. A test that shows valid differences is fair. A test that shows invalid differences is not fair. 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 questions sample only relevant knowledge and skills.
- 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.

History of fairness review. This edition of the *Fairness Review Guidelines* owes much to its predecessor documents and to the more than 30 years of experience that ETS staff have had in reviewing materials to help ensure that they are fair for all people.

- 1960s. In the late 1960s ETS test developers were encouraged - but not required - to obtain *minority reviews* of tests before they were published. There were no written criteria and no formal documentation. A test developer would ask a colleague who was a member of a minority group to review a test to determine whether any items were offensive or unfair. Changes were made at the discretion of the test developer.
- 1970s. In the 1970s the review process was enhanced by the use of written guidelines, and expanded to include women. The *ETS Guidelines for Testing Minorities* was distributed to staff in 1974, and the *ETS Guidelines for Sex Fairness in Tests and Testing Programs* followed in 1976. External documents, such as an early version of the McGraw-Hill Book Company *Guidelines for Bias-Free Publishing* and the “Guidelines for Nonsexist Language” from the American Psychological Association, were also circulated among test developers. The reviews remained voluntary and undocumented, however.

- 1980s. In 1980 the reviews became mandatory, standardized, and documented with the publication of the *ETS Test Sensitivity Review Process: Guidelines and Procedures*. The reviews became the responsibility of all ETS test developers rather than being limited to minority staff. Rigorous training in performing “sensitivity reviews” and strict adherence to the documented guidelines for writing fair items were required of all test developers. (Those who did not feel comfortable performing the reviews were not required to act as reviewers, however.) If a test developer did not agree with a sensitivity reviewer’s suggestions for change, a documented procedure for resolving disputes had to be followed.

In 1984 the sensitivity review process was expanded to include all testing program publications such as test bulletins and score interpretation guides. In 1986 the *ETS Test Sensitivity Review Process* document was revised to include people with disabilities. In 1988, the review process was again expanded to include research reports and statistical reports.

- 1990s. The 1992 edition of the *ETS Test Sensitivity Review Process* included guidelines concerning older people. In 1994 the Supplementary DIF Guidelines were added to the review process. They prohibited content, not required for valid measurement, that tended to cause inordinate differences between people in different groups who had been matched in terms of the knowledge measured by the test. In 1996 guidelines were added concerning people who are bisexual, gay, lesbian, or transgendered.

The current revision was undertaken because, after more than 20 years of use, the *ETS Test Sensitivity Review Process* document had become worn. The piecemeal addition of appendices over time to deal with new issues had resulted in more of a “patchwork” text than in an integrated whole. There was excessive redundancy because older parts were not rewritten as newer parts were added. Some sections of the document seemed outmoded because they devoted a great deal of text to issues that were current in the late 1970s, but that were no longer major concerns. The document did not adequately cover international testing and K-12 testing. Finally, the link between fairness and validity needed to be strengthened.

As societal views of fairness have evolved and as knowledge of measurement has increased, the guidelines have become more inclusive and comprehensive. The goal of the reviews, however, has remained unchanged since their informal beginnings in the 1960s – to make ETS materials as fair as possible.

Stanford von Mayrhauser
Senior Vice President and General Counsel
Educational Testing Service

INTRODUCTION

Purpose. The primary purposes of the *Fairness Review Guidelines* are to help ETS staff develop fair and valid assessments and other products and to help staff communicate appropriately with diverse groups of people. The guidelines² require the use of inclusive and unbiased language and images.

Application to ETS materials. Even though much of the content of the guidelines refers to tests, the guidelines apply to all ETS materials, in any medium, that contain language or images.

ETS requires a formal, documented Fairness Review by specially trained staff for compliance with the guidelines for all assessments³ and for all other materials intended for use by more than 50 people outside of ETS. Such materials may include test bulletins, sample items (test questions), research reports, surveys and interview protocols, Web pages, computerized and traditional instructional materials, videotapes, brochures, newsletters, form letters, proposals, and advertisements.

ETS staff who write, review, or edit assessments or materials that require Fairness Review should be trained in the application of the guidelines. Even if formal reviews are not required, it is a good practice to obtain informal reviews of materials that deal with sensitive issues.

Rationale. Compliance with the guidelines is important for effective communication and valid testing. With respect to communications, the guidelines are intended to help ensure that diverse audiences will understand the materials and not be offended by them. With respect to assessments, the guidelines are intended to help ensure that only construct-relevant factors affect test takers' scores. (Something that is construct-relevant is part of the knowledge, skills, abilities, or other characteristics a test is supposed to be measuring.) Test items that cause group differences because of construct-irrelevant factors are not fair. Test items that cause group differences because of valid factors are fair, however. It is fair to measure valid knowledge, even if the knowledge is not equally distributed across groups.

If construct-irrelevant knowledge is required to answer items in a test and the knowledge is not universally available to test takers, the validity and fairness of the test are reduced. For example, the inclusion of unnecessarily difficult reading material in the word problems of a mathematics test would be invalid and unfair to test takers with limited facility in English.

² If the word *guidelines* begins with a lowercase letter, it refers to the rules for achieving fairness contained in this document.

³ Additional guidelines apply to K-12 state tests and NAEP items. See the document *Special Requirements for Fairness Review of K-12 State Tests and NAEP Items*, published by ETS, 2003.

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 difficulty. Test takers may be distracted if they believe that a test advocates positions counter to their beliefs. Test takers may respond emotionally rather than logically to needlessly controversial material. The inclusion of such material may also have adverse effects on performance on subsequent items. Even if performance is not directly affected, the inclusion of inappropriate content or images may decrease the confidence of test takers and others in the fairness of ETS products.

Adherence to the guidelines will help to improve validity by eliminating some potentially construct-irrelevant aspects of assessments. Therefore, compliance with the guidelines is mandated by ETS policy and is required by the *ETS Standards for Quality and Fairness* (ETS, 2002).

Application to groups. The groups of primary concern for these guidelines are defined by age, disability, ethnicity, gender, national origin, race, religion, and sexual orientation.

It is possible to construct scenarios in which almost any content would be problematic for some group of people defined by special circumstances. For example, it is possible to hypothesize that a biology passage about the function of the heart might upset test takers who have a family member with heart disease. The guidelines should not be extended to prohibit such otherwise generally acceptable content.

Application to types of tests. In the application of the guidelines, it is important to distinguish between tests of specific subject-matter knowledge (content tests), and tests of general skills and abilities (skills tests). Skills tests are designed to assess a general skill, such as reading comprehension, writing, mathematical reasoning, or problem solving, which can be applied across subject-matter areas. Content tests are designed primarily to assess knowledge in a specific discipline, such as art, biology, economics, English literature, American history, nursing, or psychology.

The important distinction between content and skills tests with respect to the guidelines is that particular topics are rarely required for a skills test. Even though skills tests may have requirements for materials within broad categories such as business, social science, natural science, or humanities, there are seldom requirements for specific topics within those categories. Therefore, skills tests would seldom require material out of compliance with the guidelines for valid measurement. Content tests, however, must include content that is required for valid measurement even if it would otherwise be out of compliance with the guidelines.

Some mixed tests contain both content aspects and skills aspects. Treat the content aspects as though they were content tests and the skills aspects as though they were skills tests.

Application to domestic and international populations. The guidelines apply to materials intended for use primarily in the United States, to materials intended for mixed domestic and international use, and to materials intended for general use in many different countries outside of the United States.

Materials designed for use in a specific country other than the United States, however, will very likely require revision to one or more of the guidelines. The needed revisions will vary depending on the country for which the materials are designed. For example, a test made for use in Nigeria and a test made for use in India would require different guidelines for the representation of diversity. A test made for use in Denmark and a test made for use in Bahrain would require different guidelines for what is considered offensive or controversial. Consult the Center for Fairness in Assessment in the Office of the General Counsel concerning such revisions.

Link to validity. The purpose of the guidelines is to enhance validity, not diminish it. Nothing in the *Fairness Review Guidelines* is intended to interfere with valid measurement. Therefore, material required for valid measurement is acceptable even if it includes topics, ideas, attitudes, images, or other content that the guidelines would otherwise prohibit. For example, a detailed description of the effects of a car accident may be acceptable in a content test for licensing emergency medical technicians, even though it would be unacceptable in a skills test of reading ability.

Fairness at test design. Fairness should be addressed during the design and development phases of test creation. ETS staff and any external collaborators who write or revise test specifications should do so in light of the guidelines. Content or images that would otherwise violate the guidelines should be included in a test only if required for validity. The departures from the guidelines should be as small as is consistent with valid measurement. The required material should be documented in the test specifications. Both ETS and outside item writers should be familiar with the guidelines and the specifications before they select stimulus materials or write items.

Some programs allow source materials to be edited, if necessary, to be in compliance with the guidelines. For example, in such programs it would be permissible to delete a sentence in a reading passage, not required for comprehension of the material, if it contained a demeaning stereotype. Other programs do not allow the editing of source materials. Test developers in such programs would have to seek an excerpt that did not contain any problematic material. Programs should decide at the design stage of test creation whether or not to allow editing of source material. If editing is allowed, programs should notify test takers that source materials may have been adapted for use in the test.

Overview. This document includes six guidelines concerning the fairness of content and images in ETS materials:

- Guideline 1. Treat people with respect in ETS materials.
- Guideline 2. Minimize the effects of construct-irrelevant knowledge or skills.
- Guideline 3. Avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting.
- Guideline 4. Use appropriate terminology to refer to people.
- Guideline 5. Avoid stereotypes.
- Guideline 6. Represent diversity in depictions of people.

Following the guidelines, general procedures are described for implementing fairness reviews and resolving disputes.

GUIDELINES

Guideline 1. Treat people with respect in ETS materials.

Language and images. Use language and images that show respect for people in different groups. Avoid language and images that are belittling, condescending, contemptuous, derogatory, demeaning, disrespectful, exclusionary, insulting, or patronizing, unless such content is required for valid measurement. A history test, for example, may appropriately include material that would otherwise be out of compliance with the guidelines to illustrate certain derogatory attitudes commonly held in the past.

Societal roles. ETS materials should demonstrate that people in different groups are found in a wide range of societal roles and contexts. Do not over-represent members of any group in examples of lower-status occupations or in examples of inappropriate, foolish, unethical, or criminal behavior.

Problems and beliefs. Do not treat the problems of any of the groups of primary concern for fairness as humorous or inconsequential. Do not treat the strongly held beliefs of any of those groups as bizarre or nonsensical. It may be appropriate in some instances to indicate that there is no evidence supporting a group's beliefs, or that other people hold opposing views, but an impartial tone should be maintained unless a different tone is necessary for valid measurement.

Ethnocentrism. Another aspect of treating people with respect is avoiding any indication that any particular group is superior to other groups, or is the standard against which all other groups are measured.⁴ For example, the phrase *culturally deprived* implies that the majority culture is superior and that any differences from it result in deprivation. Avoid language and images that suggest that all people in higher-status positions are members of any single group.

Do not use language that assumes all test takers are citizens of the United States. For example, unless the context makes it clear, do not use the phrase *our country* to refer to the United States of America. Do not assume that all test takers share the point of view that the United States has taken on international controversies. For example, unless the context makes it clear, a reference to "allies" should not automatically be used to mean allies of the United States.

The word *America* should not be used without an explanatory context to refer only to the United States. Similarly, the term *government* should not be used without explanation to refer particularly to the United States government, the *South* should not be used without explanation to refer to the southern tier of states in the United States, and so forth.

⁴ This is not intended to prohibit the use of reference groups in statistical analyses.

Underlying assumptions. An underlying assumption is an unstated proposition on which an item or other material is based. Avoid content or images based on negative underlying assumptions about groups of people. Avoid inappropriate underlying assumptions that would violate the guidelines if they had been stated. For example, consider the sentence, “The pioneers and their wives crossed the plains.” The sentence requires the inappropriate underlying assumption that only men were pioneers.

In general, avoid using *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.

Guideline 2. Minimize the effects of construct-irrelevant knowledge or skills.

Construct-irrelevant difficulty. If construct-irrelevant knowledge or skill is required to answer an item, and the knowledge or skill is not equally distributed, then the item is not fair. The following types of knowledge and skill have caused problems with construct-irrelevant difficulty in some situations. They should be avoided unless they are clearly construct-relevant. A recurrent theme is the avoidance of specialized vocabulary unless it is required for valid measurement. What is considered unnecessarily specialized will vary with the age and sophistication of the test takers and with what is required for valid measurement.

- **Charts, maps, graphs, and other visual stimuli.** Avoid the use of charts, maps, graphs, and other visual stimuli if they are arbitrarily chosen as one of many possible means of testing some particular point. Test takers may lack skills in reading charts, maps, and graphs and still have mastered the construct that an item is intended to measure. Furthermore, the unnecessary use of visual stimuli adds construct-irrelevant difficulty for blind or visually impaired test takers. Such stimuli are difficult to reproduce in Braille or to describe orally. Charts, maps, graphs, and other visual stimuli are acceptable if the ability to read and understand them is part of the construct, or one of the more common means of presenting data in the area being tested.
- **Unnecessarily difficult words, figures of speech, idioms, or syntactic structures.** Avoid unnecessarily difficult language unless the purpose of the item is to measure the ability to deal with such language. What is considered unnecessarily difficult will vary with the age and sophistication of the test takers. Use the simplest and most straightforward language consistent with valid measurement to help avoid construct-irrelevant difficulty for test takers who are not native speakers of English, and for test takers who are deaf or hard of hearing. 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. Complicated syntax may be appropriate if the purpose of the test is to measure the ability to read difficult material.

- **Elitism.** Avoid the use of words or topics generally associated with wealthier social classes, such as *penthouse*, *polo*, *regatta*, *yacht*, and the like, unless required for valid measurement. Avoid unnecessary depictions of people spending large amounts of money on luxuries. Avoid uncommon financial vocabulary, such as *arbitrage*, *equities*, *estate planning*, *junk bonds*, *stock options*, *tax-free bonds*, *venture capital*, etc., unless such terms are required for valid measurement or related to the purpose of the test.
- **Specialized farm-related words.** Avoid requiring knowledge of uncommon farm-related words, such as *combine* and *thresher*, unless required for valid measurement. Common words like *field*, *harvest*, and *plow* are acceptable.
- **Specialized legal words.** Avoid requiring knowledge of uncommon legal terms when the goal is to measure some other skill or knowledge. Words such as *judge* and *jury* are acceptable because they are common knowledge. Words such as *subpoena* or *tort* would not be acceptable unless the goal is to measure knowledge of legal terminology.
- **Military topics.** Avoid items with a primary focus on military topics, such as armed forces, battles, military strategy, wars, weapons, etc., unless required for valid measurement. Avoid unnecessary use of specialized words related to weapons, such as *rapier*, *mortar*, or *breech*. Avoid requiring knowledge of the functions of parts of weapons or how weapons work unless that is the intended point of measurement.

Mention of war, weapons, or any other military topic in a stimulus primarily concerned with some other topic is acceptable. For example, a passage about Joan of Arc that mentioned that she wore armor and carried a sword would be acceptable. Items on military topics are acceptable whenever required for valid measurement, as in some history tests.

- **Specialized political words.** Avoid requiring knowledge of uncommon words related to politics unless required for valid measurement. Common words such as *president* or *vote* are acceptable. Terms such as *alderman* or *pork barrel* are not acceptable if the purpose of the item is to measure skills other than knowledge of political terminology.
- **Regionalisms.** Unless required for valid measurement, avoid words, phrases, and concepts more likely to be known by people in some regions than in others. When there is a choice, use generic words rather than their regional equivalents. For example, more test takers - particularly those outside of the United States - are likely to understand the generic word *sandwich* than are likely to understand the regionalisms *grinder*, *hero*, *hoagie*, or *submarine*. Names used for political jurisdictions, such as *borough*, *province*, *county*, *parish*, etc., vary greatly across regions. Knowledge of their meaning should not be required to answer an item unless such knowledge is part of the construct.

- **Religious knowledge.** ETS tests are taken by people of many different religions as well as by people of no religion. Do not require specific knowledge about any religion to answer an item unless such knowledge is part of the construct. If the knowledge is part of the construct, take care to use only the information about religion required 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.
- **Scientific and technical words.** Even though the specifications of many verbal skills tests call for items in the content category *science*, the goal is to measure verbal ability rather than knowledge of scientific subject matter. Therefore, avoid requiring knowledge of specialized words unless required for valid measurement. Common words such as *computer*, *microscope*, *thermometer*, and *degree* are acceptable. More specialized terms, such as *byte*, *lumen*, and *vacuole*, are not acceptable if the goal is to measure general vocabulary. Any scientific and technical words required for valid measurement are acceptable.
- **Spatial skills.** Avoid the measurement of spatial skills (visualizing how objects or parts of objects relate to each other in space) unless the measurement of such skills is the purpose of the item. The unnecessary measurement of spatial skills adds construct-irrelevant difficulty for blind or visually impaired test takers and may be a source of invalid differences between groups.
- **Sports.** If knowledge of a sport is needed to answer an item, the item should not appear in a general skills test. Even if knowledge of the sport is not needed to answer an item, avoid unnecessary references to particular sports in the setting of an item. It is acceptable to use general words associated with sports such as *game* and *score*. A passage that refers to a sport but does not focus on the sport is acceptable.
- **Specialized words associated with tools and machinery.** Avoid requiring knowledge of specialized words associated with tools and machinery, such as *torque*, *flange*, and *chuck*, unless required for valid measurement. Avoid requiring knowledge of what tools do, such as the purpose of a die, a compressor, or a router, unless such knowledge is required for valid measurement. 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 required for valid measurement.
- **Specialized words associated with transportation.** Avoid requiring knowledge of specialized vehicles, such as ketch or biplane, or of parts of vehicles (e.g., strut, cam, boom) unless it is required for valid measurement. General words such as *train*, *boat*, and *motor* are acceptable.

- **United States culture.** ETS tests are often taken by people in many different countries who may not be familiar with United States culture. For such tests, do not require specific knowledge of United States culture to obtain the answer to an item unless the item is supposed to measure such knowledge. For example, do not require knowledge of American 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, entertainers, geography, history, holidays, institutions, laws, measurement systems (inches, pounds, quarts, degrees Fahrenheit, etc.), plants, politicians, political systems, slang, sports, television shows, or wildlife specific to the United States.

Guideline 3. Avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting.

Problematic topics. Material that is unnecessarily controversial, inflammatory, offensive, or upsetting may serve as a source of construct-irrelevant difficulty and should be avoided whenever possible. Some reasonably controversial material may be necessary for valid measurement, however, 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.

As noted in the Introduction, material required for valid measurement is acceptable even if it might otherwise be out of compliance with the guidelines. If such material must be used, it should be handled in a conscientious, balanced, sensitive, and objective manner. Furthermore, the least controversial, inflammatory, offensive, or upsetting material that will meet the requirements should be used.

Make clear to test takers that such materials are used to assess knowledge or skill and do 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 by stating that the material does not represent the views of ETS or the testing program.

What is considered excessively controversial, inflammatory, offensive, or upsetting will vary. Material that is neutral or pleasing to some groups may be offensive to other groups. Material that is acceptable for graduate students may be inappropriate for children in elementary school. Material that is not offensive in the United States may be offensive in some other countries. Furthermore, the way some topics are treated must be taken into account in determining whether or not the material is acceptable.

Topics to avoid. Some topics are so sensitive that the best strategy is to avoid them if it is possible to do so. Any list of topics to avoid can only be illustrative rather than exhaustive. A topic is not necessarily acceptable merely because it has not been included

on this list. Avoid language or images that deal with topics as controversial, inflammatory, offensive, or upsetting as the following unless required for valid measurement or the purpose of the communication.

- Abortion
- Abuse of people or animals
- Euthanasia
- Experimentation on human beings or animals that is painful or harmful
- Genocide
- Human sexual behavior, including contraception
- Hunting or trapping for sport
- Rape
- Satanism
- Torture
- Witchcraft

Topics requiring extreme care. There are topics that need not be avoided entirely, but that must be handled in a very sensitive manner. Treat topics that are as controversial, inflammatory, offensive, or upsetting as the following in ways that limit their problematic aspects unless a different treatment is required for valid measurement or the purpose of the communication. It is a good practice to obtain a preliminary fairness review of sensitive materials before time is expended on them.

- Accidents, illnesses, or natural disasters. Avoid dwelling on gruesome, horrible, or shocking aspects of accidents, illnesses, or natural disasters unless required for valid measurement. Other aspects of those topics are acceptable. For example, it is acceptable to address the prevention of accidents, the causes of illness, or the occurrences of natural disasters.
- Death and dying. Do not focus on gruesome details associated with death and dying unless required 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. It is undeniable that evolution is controversial, so it is preferable not to focus on the topic. It is also undeniable, however, that evolution is a core concept in biological science. Furthermore, topics associated with evolution are important in several other disciplines. The most sensitive aspect of evolution appears to be the evolution of human beings. Therefore, for skills tests other than K-12 state tests, any aspect of evolution except the evolution of human beings is allowed. Any aspect of evolution is allowed on content tests as required for valid measurement. (See the document *Special Requirements for Fairness Review of K-12 State Tests and NAEP Items*, published by ETS in 2003, for the treatment of evolution in K-12 state tests.)
- Religion. It is preferable not to focus on religion in ETS materials unless required for valid measurement. If required, materials about religion should be as objective as possible. Do not support or oppose religion in general or any specific religion in ETS materials. Do not praise or ridicule the practices of any religion.

- Slavery. Except as required 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 would be acceptable even if it mentioned slavery. A passage on the abolitionist movement would be acceptable even if the passage touched on slavery. A discussion of the economic value of slaves would not be acceptable in a skills test, however.
- Suicide or 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 in a skills test or to glorify suicidal behavior. Similarly, details of substance abuse should be avoided unless required for valid measurement.
- Violence and suffering. Do not focus on violent actions, on the detailed effects of violence, or on suffering unless required for valid measurement. Violence and suffering are too pervasive in art, biology, history, literature, and in most aspects of human and animal life to exclude them completely from all materials, even in skills tests. Do not, however, dwell unnecessarily on the gruesome, shocking aspects of violence and suffering. For example, it is acceptable to discuss the food chain even though animals are depicted eating other animals. It would not be acceptable, however, to include a vivid description of wolves disemboweling a struggling fawn unless required for valid measurement.

Materials 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. There are many culture-specific taboos. For example, people in some cultures consider it offensive to use the left hand when offering an object to another person, while people in some other cultures consider it offensive to use only one hand for the same task. It is probably impossible to avoid all such taboos. However, as a general rule, if some aspect of any content or image is believed to be offensive or confusing to some group of test takers, it should be avoided if

- the aspect is not required for valid measurement,
- the aspect is not inordinately difficult to eliminate, and
- the elimination of the aspect will not offend or confuse other test takers.

For materials designed for use by people in many different countries, avoid the following types of images or descriptions unless required for valid measurement or the purpose of the communication.

- 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.

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.⁵

Advocacy. Do not use test content to advocate any particular cause, lifestyle, or ideology unless required for valid measurement. Items and stimulus materials should be neutral and balanced whenever possible. Do not “take sides” on any controversial issue unless required for valid measurement. Test takers who have opposing views may be disadvantaged by the need to overcome their beliefs to respond to items in accordance with the point of view taken in the stimulus material.

Some types of items require the presentation of a particular point of view, as when an argument is to be evaluated. Such items should be no more controversial than is required for valid measurement.

Requirements for research reports. Some research may necessarily focus on controversial and/or inflammatory issues. In such research, use language that is emotionally neutral to discuss the issues. Avoid gratuitous controversial statements or examples. Material that may have a negative emotional impact on a subgroup is acceptable in research reports only if it is necessary to the research. Present the material in a way that will reduce such impact.

Guideline 4. Use appropriate terminology.

Appropriate terminology. Do not attach unnecessary labels to people. If a person's membership in a group is not relevant, do not mention it.

If group identification is necessary, it is generally most appropriate to use the terminology that group members prefer. Explicitly derogatory names for groups should not be used in ETS materials, even if they are used by some group members.

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 *Asian American* are not hyphenated.)

⁵ Some of the guidelines may be in conflict with the desire to meet certain cultural expectations. For example, the requirement for gender balance may not meet cultural expectations in some countries. Compliance with the guidelines is required, however, unless revisions have been made for tests designed for use in a particular country.

So-called minority groups are becoming the majority in many locations in the United States, and are the majority in many other countries. Therefore, although the terms are still acceptable, try to reduce the use of *minority* and *majority* to refer to groups of people.

Discussions of appropriate terminology for various population groups follow⁶.

- **Appropriate terminology for people with disabilities.** In the first reference to someone with a disability, focus on the person rather than the 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*). These 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*.

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 elevator.

Avoid euphemistic or patronizing terms such as *special, physically challenged, and inconvenienced*. Avoid the use of such words and phrases as *inspirational, courageous, achieving success in spite of a disability, and overcoming a disability*.

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.

The following commentary discusses some of the terms commonly used to describe people with disabilities.

Abnormal, Normal	Unacceptable. The terms <i>normal</i> and <i>abnormal</i> are not appropriate for referring to people except in medical contexts.
Blind	The noun form, <i>the blind</i> , is not acceptable except in the names of organizations or in historical material. The adjectival form, <i>a blind person</i> , is acceptable in subsequent references if <i>person who is blind</i> is used in the initial reference.

⁶ For K-12 state or district tests, determine the client's preferences concerning terminology.

Deaf	Acceptable as an adjective, but sometimes the term <i>deaf</i> or <i>hard of hearing</i> is used as a noun, e.g., School for the Deaf. References to the Deaf community should be capitalized when referring to the group, but references to a person can be lowercased. Avoid the phrases <i>deaf and dumb</i> and <i>deaf mute</i> .
Down syndrome	Use the term <i>Down syndrome</i> rather than <i>Down's syndrome</i> . Avoid the obsolete and inappropriate term <i>Mongoloid</i> .
Hearing impaired	The Deaf community prefers <i>deaf and hard of hearing</i> to cover all gradations of hearing impairment.
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 <i>person with a learning disability</i> .
Mentally ill	Generally unacceptable. The preferred term is a <i>person with a psychological or emotional disability</i> .
Mentally retarded	Acceptable, but preferred usage is a <i>person with mental retardation</i> . Can also say, <i>developmentally disabled</i> or <i>developmentally delayed</i> . Do not use the term <i>retarded</i> by itself.
Paraplegic, quadriplegic	Acceptable as adjectives, not as nouns.
Physical disability	Acceptable.
Spastic	Unacceptable to describe a person. Muscles are spastic, not people.

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.

- **Appropriate terminology for African American people.** The terms *Black and African American* are both acceptable. Note that *Black* should begin with an uppercase letter. The terms *Negro* and *Colored* are not acceptable except when embedded in clearly historical contexts as required for valid measurement, 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*.)

- **Appropriate terminology for Asian American people.** The terms *Asian American*, *Pacific Island American*, and *Asian/Pacific Island American* should be used as appropriate. If possible, use specific terminology such as *Bangladeshi American*, *Chinese American*, *Filipino American*, *Japanese American*, and so forth. Do not use the word *Oriental* to describe people.
- **Appropriate terminology for Hispanic American people.** The terms *Hispanic American* and *Latino American* are both acceptable and may be used as appropriate. For women, use the term *Latina* in place of *Latino*. 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 the words. Where possible, use a specific group name such as *Cuban American*, *Dominican American*, *Mexican American*, and so forth as appropriate.
- **Appropriate terminology for Native American people.** The terms *American Indian* and *Native American* are both acceptable. Avoid use of the term *Eskimo* for people who are more acceptably called *Alaskan Natives* or *Inuit*. 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 more specific group names they use for themselves. However, this 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*.
- **Appropriate terminology for 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). The term *European American* may also be used. The term is gaining some currency because of its parallelism with *African American*, *Asian American*, etc.
- **Appropriate terminology for 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. *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*.

People of one sex must not be described by physical attributes when people of the other sex are described by mental attributes or professional position. Irrelevant references to a person's appearance or attractiveness are not acceptable.

Women should generally be referred to as *women*. If they are 18 or older, they should not be referred to as *girls*. Men should generally be referred to as *men*. If they are 18 or older, they should not be referred to as *boys*.

Language that assumes that all members of a profession are members of one sex 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. Role labels such as *scientist*, *executive*, and so forth include both men and women. Do not use expressions such as *the soldiers and their wives* that assume only men fill those 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.

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 material. Alternating generic *he* and generic *she* is unacceptable because neither word should be used to refer to all people. Avoid *he/she* and *(s)he*. Examples of unacceptable usages of *man* and *he* with acceptable alternatives follow.

Unacceptable

Acceptable

mankind, man

humanity, human beings, people

manmade

synthetic, artificial

manpower

workers, personnel, labor, work force

fireman, mailman,
salesman, insurance
man, foreman

firefighter, mail carrier, sales representative, insurance
agent, supervisor

chairman

chair, presiding officer, leader, moderator
(*chairman* and *chairwoman* are acceptable when
referring to specific men and women. Do not use
chairman for a man and *chairperson* for a woman
because the terms are not parallel.)

If a student studies, he
will learn.

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.

- **Appropriate terminology for people who are bisexual, gay, lesbian, or transgendered.** Issues of human sexuality should be avoided unless required for valid measurement, as indicated in Guideline 3. Therefore, identify people by sexual orientation only when it is construct-relevant to do so. Do not use the labels gratuitously.

The words *bisexual*, *gay*, *lesbian*, and *transgendered* are all acceptable. Because some people assume that *gay* refers to men only, use *gay* or *gay people* only when prior reference has specified the gender composition of this term.

Avoid using the term *homosexual* outside of a scientific or historical context. Do not use the term *queer* to refer to sexual orientation, even though the term has some currency among gay and lesbian political activists and scholars. Use the phrase *sexual orientation* rather than *sexual preference*.

- **Appropriate terminology for older people.** It is best to refer to older people by specific ages, for example, *people age 65 to 75*. It is also acceptable to refer to *older people*. Avoid *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.

Guideline 5. Avoid stereotypes.

Definition of stereotype. A stereotype is a conventional, overgeneralized, and oversimplified conception of the characteristics of a group of people. Stereotypes attribute characteristics to groups on the basis of age, disability, ethnicity, gender, national origin, race, religion, or sexual orientation. Stereotypes ignore differences among members of the group. Generally, any statement or image that ascribes the same characteristic to all, or even most, members of a group (unless the group was composed on the basis of that characteristic) is a stereotype. Do not imply that all members of a group share the same culture, as opposed to recognizing substantial cultural variations within groups. Avoid stereotypes in language and images unless required for valid measurement, or necessary for research. If stereotypes are required, use the least offensive stereotypes that will result in valid measurement.

The terms *stereotypical* and *traditional* overlap in meaning but are not synonymous. Depicting an individual engaged in a traditional activity (such as a woman cooking) does not necessarily constitute stereotyping. As long as the material as a whole does not depict members of a group engaged exclusively in traditional activities, such references are acceptable. However, because some computerized tests are assembled as they are being administered, it is often impossible to know whether group members shown in traditional roles will be balanced by group members shown in nontraditional roles. Therefore, in items intended for use in such computerized tests, avoid showing group members in traditional roles, if such roles border closely on stereotypes.

Stereotypes. Avoid using phrases that encapsulate stereotypes such as *Dutch uncle*, *Indian giver*, *run like a girl*, *women's work*, *man-sized job*, and so forth. In addition, no group defined by age, disability, ethnicity, gender, national origin, race, religion, or sexual orientation should be stereotyped as superior or inferior to any other such group with regard to any of the following, or similar, characteristics.

- Contribution to society
- Dependence on welfare
- Dialect or language usage
- Generosity
- Honesty
- Impulsiveness
- Industriousness
- Leadership ability
- Morality
- Physical appearance
- Quality of culture

Special requirements for tests. Unless the testing of stereotypes is specifically required for valid measurement, avoid stereotypes in tests even 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.

Passages that explore the process of stereotyping, without including negative stereotypes, are acceptable in skills tests. Content tests may include stereotypes to the extent required for valid measurement.

Special requirements for research. To avoid the reinforcement of stereotypes in discussions of group differences in research and statistical documents, do not imply that there is a biological or social cause not demonstrated by the research. Do not imply that there are innate differences between groups when differences between environments may explain the result. Do not imply that the responses of two groups fall at opposite ends of a single continuum when this is not accurate.

Avoid generalizations about groups unless supported by data. Be specific about sources of variation in group differences that are reported. Do not interpret group differences without appropriate statistical comparisons, consideration of the magnitude of the effect, or consideration of the degree of overlap between distributions. When presenting results of research about groups, describe the situational context and the environment in which the research was done. If statements about a group have been demonstrated in previous research, cite the previous research. Be careful about attributing causality to group membership based only on correlational data. If such attributions are made, specify the correlational nature of the relationship.

Guideline 6. Represent diversity in depictions of people.

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 Albert Einstein in one item is not balanced by a mention of a placeholder female name such as Jane in another item.

The gender balance of content tests should be appropriate to the content area. In occupational tests, do not depart greatly from the gender distribution of the members of the occupation. For example, showing half of the nurses in a test as male or half of the automobile mechanics as female would be unrealistic and possibly disconcerting to test takers⁷. At least 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. The ideal racial and ethnic balance in a test would reflect the diversity of the test-taking population. It is not feasible, however, to show members of every group in the test-taking population in every test. As a reasonable compromise, in skills tests made for use primarily in the United States, strive to have about 20 percent of the items that mention people represent African American people, Asian American people, Latino American people, and/or Native American people. Try to depict several different groups in a test.⁸

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 placeholder names commonly associated with various groups (e.g., Juan, Kimani, Latisha, Matsuko, Ram). Do not add unnecessarily to the linguistic loading of the item by using names that are inordinately difficult for test takers to decode.

For skills tests with mixed domestic and international populations and for skills tests made for general international use, the representation of diversity among 20 percent of the items that mention people may include African people, Asian people (including Indian people), Latino people, and such indigenous peoples outside of the United States as the Guarani Indians of Paraguay.

⁷ Reflecting the gender balance of the occupation may not be appropriate if only small numbers of people are depicted. If only two nurses are mentioned, for example, depicting one of the nurses (50 percent) as male would not reflect the gender balance of the occupation. In such situations, however, represent both genders because it would not be unusual for small groups to have a gender balance that did not reflect the gender balance of the occupation.

⁸ Some programs may need time to comply with the increase in the required proportions. Such programs should discuss their needs with the Center for Fairness in Assessment in the Office of the General Counsel. For adaptive tests, the 20 percent representation should be assured at the pool level.

For skills tests with mixed domestic and international populations, at least one of the groups represented should be from the United States. For skills tests made for general international use, it is acceptable, but not required, to include any of the United States groups listed above. Tests made for specific countries should reflect the test-taking populations in those countries.

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 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, Jay's 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 social workers as White and all clients as Black or Hispanic.

GENERAL PROCEDURES ⁹

Essential Aspects of Fairness Review

ETS makes many different types of tests and other materials subject to fairness review, under many different types of circumstances, in several locations. Some variation in Fairness Review procedures is to be expected. However, certain basic elements must be maintained in all Fairness Reviews.

- The Fairness Reviewer must have been trained in Fairness Review or have had the original training updated within the last five years.
- The Fairness Reviewer should have no stake in the test or other material being reviewed. (This may not be possible for some materials, such as those in languages for which few speakers are trained as Fairness Reviewers.)
- People who submit materials for review should not be able to select the particular Fairness Reviewers who will review their materials.
- The Fairness Review must be done with respect to the most recent version of the *Fairness Review Guidelines*.
- For tests, the Fairness Reviewer must have access to the test specifications and be aware of the characteristics of the test-taking population. The Reviewer should have access to all components of the test that a test taker would have, such as audiotapes (or scripts) and visual materials, in addition to the items. For materials other than tests, the Reviewer should know the purposes of the materials and the nature of the people for whom they are intended.
- The Fairness Review, including tallying for race and gender balance when completed tests or other materials are being reviewed, must be documented. The Reviewer should distinguish between comments for which action is optional and Fairness Review Challenges that require a response.
- If a Challenge is raised, the Fairness Reviewer must cite the specific guideline(s) that have been violated.
- Items or materials that have been challenged cannot be released outside of ETS until the Challenge has been resolved. The resolution and the Fairness Reviewer's acceptance of the resolution must be documented. If resolution cannot be reached, the disputing parties should follow the dispute-resolution procedures described below.
- If an Editor believes that an item or other material that had passed Fairness Review is not in compliance with the guidelines, the Editor should raise the issue with the original Fairness Reviewer. If the Fairness Reviewer agrees, the item or other material should be treated as though the Fairness Reviewer had made the Fairness

⁹ For detailed procedures for implementing Fairness Reviews in the Test Creation System, for research reports, and for publications and other materials, see the document *Procedures For Application of the ETS Fairness Review Guidelines* published by ETS, 2003.

Challenge. If the Fairness Reviewer does not agree, the item or other material moves on unless the Editor wishes to further dispute the judgment of the Fairness Reviewer. In that case, the Editor should follow the dispute-resolution procedures described below.

Preliminary Review

Potentially controversial material, or material that is very expensive to change at later stages, should receive a Preliminary Fairness Review before any substantive work is done. For example, scripts and story boards for a videotaped presentation should be reviewed before actors are hired, rehearsals are held, and taping and editing are done. The Preliminary Review is optional, but it is strongly recommended as a way to reduce the risk of expending resources on materials that later will be found to be out of compliance with the guidelines. The Preliminary Review should be as rigorous as is a Final Review, and be documented. The Preliminary Reviewer must meet all of the qualifications of the final Fairness Reviewer listed above.

The normal Fairness Review at the end of the process remains mandatory, even if a Preliminary Review had been obtained, because the Preliminary Review is necessarily incomplete. If the Final Reviewer wishes to raise a Fairness Challenge on material previously cleared by the Preliminary Reviewer, he or she should try to reach agreement with the Preliminary Reviewer and the test developer. If the parties cannot reach agreement after discussing the issue, they should follow the dispute-resolution procedures described below.

Resolution of Disputes

Judgment is required in the application of the guidelines. It is impossible to develop rules and examples that will cover every situation. In some cases, application of the guidelines will require the careful evaluation of competing priorities. Therefore, equally reasonable and informed people may occasionally reach different decisions about the appropriate action to take. If the parties to a dispute about application of the guidelines cannot reach agreement after discussing the issue, the following steps should be taken¹⁰:

- The disputing parties will notify the cognizant Fairness Review Coordinator, who will try to help them find a solution acceptable to both parties.
- If agreement is not reached, the Fairness Review Coordinator will notify the Center for Fairness in Assessment (CFA) in the Office of the General Counsel. CFA staff will discuss the problem with the disputing parties and the Fairness Review Coordinator, and will propose a resolution. Ordinarily, the resolution should be proposed by the end of the next working day following the discussion.

¹⁰ For disputes involving K-12 tests, see *Special Requirements for K-12 State Tests and NAEP Items* published by ETS, 2003.

- If the proposed resolution is not accepted by both parties, a three-member subset of the Fairness Review Steering Committee will meet with the disputing parties to discuss the problem and will propose a resolution. Ordinarily, the resolution should be proposed by the end of the next working day following the meeting.
- If the proposed resolution is not accepted by both parties, the General Counsel, who is the cognizant officer for Fairness Review, will make a final ruling based on written arguments provided by the disputing parties. Ordinarily, the disputing parties should complete the written arguments within two working days following the notification that they will be needed.

Need for Additional Fairness Review

Items or other materials that are in use should be reviewed for fairness at least once every five years. Items or other materials should be reviewed if they will be used for a purpose or population substantively different from those for which they were initially reviewed. If more than 25 percent of the items in a paper-based test are deleted, revised, or replaced, the new form should receive a Fairness Review.

SOME USEFUL REFERENCES ON FAIRNESS IN ASSESSMENT

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

American Psychological Association. (1977). *Guidelines for nonsexist language in APA journals*. Washington, DC: Author.

American Psychological Association. (2001). *Publication manual of the American Psychological Association*. Washington, DC: Author.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. Holland & H. Wainer, (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

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

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

Educational Testing Service (1992). *ETS test sensitivity review process: Guidelines and procedures*. Princeton, New Jersey: Author.

Flaugher, R.L. (1978). The many definitions of test bias. *American Psychologist*, 33, 671-679.

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

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

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

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

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