



# **ETS International Principles for Fairness Review of Assessments**

**A Manual for Developing  
Locally Appropriate Fairness Review  
Guidelines in Various Countries**



**ETS International Principles for  
Fairness Review of Assessments**

**A Manual for Developing Locally Appropriate  
Fairness Review Guidelines in Various Countries**

## 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.” The *ETS International Principles for Fairness Review of Assessments* supports this mission by helping to ensure that tests created under the auspices of ETS for a country other than the United States are fair for test takers in that country.

The *Principles* serves as the basis for developing appropriate guidelines for fairness review of test items for a particular country other than the United States. ETS recognizes that each country is unique and that what is considered acceptable in one country may not be suitable in another country. There are, however, principles for fairness in assessment that are applicable to every country.

Using the *Principles*, test developers in any country can generate specific, locally appropriate guidelines for fairness review that will enable them to design and build assessments that are fair for the intended test takers within the country.

I am pleased to issue the 2009 version of the *ETS International Principles for Fairness Review of Assessments*. The document will help ETS meet its mission to further education for all people worldwide.

Glenn Schroeder  
Senior Vice President and General Counsel  
Educational Testing Service

## INTRODUCTION

**Purpose:** The primary purpose of the *ETS International Principles for Fairness Review of Assessments* is to help ensure that tests made under ETS auspices for use in a specific country other than the United States will be fair and appropriate for test takers in the country in which the test will be used.<sup>1</sup>

The principles are intended to help the people who design, develop, and review 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.

**Intended Use:** ETS recognizes that what is considered fair will vary from country to country. ETS is not attempting to impose its specific fairness review guidelines, which were designed for use primarily in the United States, on other countries.<sup>2</sup>

There are, however, some principles for fairness that are applicable to every test, regardless of the country for which the test is made. For example, every test should exclude material that is unnecessarily offensive or upsetting to test takers. Even though the principle of avoiding such material is universal, exactly what is considered offensive or upsetting to test takers will vary from country to country. Therefore, specific fairness guidelines based on the general principles are needed for each country.

This manual describes the principles for fairness in assessment and gives step-by-step directions on how to supplement the principles with locally appropriate content and examples to result in clear and specific fairness review guidelines for use in a particular country.

The guidelines are designed to help people perform fairness reviews of tests. The reviews are intended to locate and remove or revise any materials that may be unfair. The reviews should be carried out by people who are familiar with the country's language(s), culture(s), customs, and instructional practices. Although the focus of this manual is on tests, it also applies to related documents such as test descriptions, practice materials, administrator's manuals and essay scoring guides.

---

<sup>1</sup> This manual is copyrighted but not confidential. ETS allows use of the manual by all who wish to enhance the fairness of their tests.

<sup>2</sup> For tests made for use in the United States and worldwide, rather than for a specific country other than the United States, please see *ETS Guidelines for Fairness Review of Assessments* (ETS, 2009).

**Regardless of the local guidelines that are set, no test made under ETS auspices should contain material that incites hatred or contempt for people on the basis of age, disability, ethnicity, gender, national or regional origin, native language, race, religion, socioeconomic status, or sexual orientation.**

**Intended Readers:** This manual is written for the people who will develop the locally appropriate fairness review guidelines. Much of the manual addresses the leader of that effort. For clarity and brevity, the leader will be addressed as “you” in the remainder of this manual.<sup>3</sup> Similarly, references to “your country” refer to the country for which the locally appropriate guidelines are being developed. The manual provides information for the leader on how to plan and run a meeting in which the local guidelines will be developed, and the manual can be used as a workbook by the participants at that meeting.

**Overview:** Following this introduction, the manual provides step-by-step directions for how to generate locally appropriate fairness review guidelines, including

- an explanation of some technical concepts;
- an introduction to the three principles for fairness review of assessments;
- sections on developing specific, locally appropriate guidelines for each of the principles;
- a section on additional guidelines that might be needed for tests designed for children;
- procedures for doing fairness reviews;
- a brief concluding section; and
- a short list of some useful references on fairness in assessment.

## HOW TO DEVELOP LOCALLY APPROPRIATE GUIDELINES

**Start Early:** Ideally, the development of specific fairness review guidelines based on the principles should take place before the test development process begins. The people who write or review test specifications, those who write and review test items, and those who assemble and review tests should all be familiar with the fairness review guidelines before they perform their tasks. It is far better to avoid the inclusion of inappropriate material in a test than it is to remove such material after it has been included. In any case, the guidelines must be completed in time for all items to be reviewed before they are administered to test takers.

It will probably take several months to complete the process of developing locally appropriate guidelines. ETS recommends pooling the opinions of a number of diverse people to help you develop the guidelines. You will be making decisions about what various groups consider offensive, for example. It is extremely helpful to have the advice of members of different groups in making such decisions. Therefore, you will need to

---

<sup>3</sup> If you would like the help of ETS in developing locally appropriate fairness review guidelines, please go to [www.ets.org](http://www.ets.org), click *Global Programs and Services*, and then click *Contact Us*. Direct your message to the United States offices. Depending on the level of help that you desire, ETS can offer consulting services or complete the task of developing the guidelines for you.

identify the people who will help you develop the guidelines, invite them to meetings, find dates and places that are convenient, and hold the meetings. The author will need at least a month to write the resulting fairness guidelines. Then time will be required to have the draft guidelines reviewed, revised, and accepted. Typically, the document will require two or three rounds of review and revision before it is considered acceptable. Finally, the people who write items, edit items, review items, or assemble tests will have to be trained to use the guidelines.

### **STEP 1: IDENTIFY AND OBTAIN STAFF**

**Decision Makers:** Determine who (or what group) has the authority to make the final decision about the fairness review guidelines to be used. In many countries, educational testing is a function of the government. In such countries, the decision maker is likely to be a government official. If testing is not controlled by the government, the decision maker may be the head of the organization that commissioned the development of the test being reviewed. In any case, it is important to determine who has the ultimate power to accept, reject, or revise the guidelines that are developed.

Read through this manual, make your plans for developing the guidelines, and have the plans approved by the decision maker(s) before you proceed.

**Panelists:** While it is possible for a well-informed individual to write fairness review guidelines, ETS believes that the task of augmenting the general principles to form specific guidelines is best accomplished by a panel that includes a diverse group of people who are very familiar with your country and who are also familiar with the population of test takers. Therefore, the manual will assume that you are using such a panel. For tests made for use in your country's schools, include teachers among the panelists because knowledge of curricula and instructional practices is important for such tests.

You should select panelists who represent the important subgroups of the country's population. For example, if there are both male and female test takers, there should be both male and female panelists. If there are several official languages, members representing each language group among the people taking the test should be included. If there are significant differences among regions of the country, then representatives from each of the regions should be included. If there are different racial, ethnic, or religious groups within the country, then members of the various groups should be included on the panel to the extent possible, and so forth.

Panels that are too large are difficult to manage, and panels that are too small are not likely to be representative of all the different groups among the test takers. Usually, a panel of seven to twelve people is a good compromise.

**Reviewers:** Identify the people who should serve as reviewers of the draft guidelines. As is the case with the panelists, the reviewers should be familiar with the country and the population of test takers and should represent important subgroups of the country's

population. For tests used in schools, include some teachers among the reviewers. The reviewers should definitely include representatives of the people who will actually use the guidelines in their test development activities.

**Chair of the Panel:** You may choose to chair the panel yourself, or you may select one of the other participants to do it. The chairperson needs good meeting-facilitation skills and should become familiar with this manual.

**Note Takers:** Ideally, one or two people who are not participating in the discussions should be available to take notes about the decisions reached by the panelists during the meeting. If such people are not available, one or two panelists should be asked to take notes of the decisions reached by the group.

**Author:** If you do not plan to write the guidelines yourself, identify the person who will be the author before the panelists begin their deliberations. Ideally, the author will be an experienced writer who is highly accomplished in the language in which the guidelines will be written and who is used to working with multiple reviewers.

The designated author should attend the meetings of the panelists. He or she may or may not be a panelist, but the author should feel free to ask panelists for further information about any of the guidelines proposed. If the author does not clearly understand the intent of the panelists, the resulting document may not reflect the panelists' wishes.

## STEP 2: PREPARE FOR THE MEETING

**Arrange for Meeting:** It is best if the panelists meet to develop fairness review guidelines because discussions among the panelists will be extremely helpful in reaching consensus on complicated issues. It is reasonable to expect the task to take about two and a half to three full days of discussion time.

To encourage discussion, it is best if the panelists are arranged in a circle, preferably around a large table. If physical meetings cannot be held, virtual conferences via the Web or teleconference facilities are acceptable.

**Distribute this Manual:** Before the meeting, distribute this manual to all of the panelists. If necessary, prepare translated copies for the panelists.<sup>4</sup> The translators should be proficient enough to convey subtle distinctions in the languages involved. For example, in American English, there is an important distinction between the seemingly very similar phrases “colored people” (which is considered unacceptable in most circumstances) and

---

<sup>4</sup> ETS grants permission to prepare and distribute, but not sell, translated versions of this document. This permission is conditioned on the author of the translation being clearly identified, and the inclusion of a disclaimer, in all copies of the translation, to the effect that ETS has not reviewed or endorsed the translation. ETS requests that a copy of the translated document, or a URL from which the document may be downloaded, be sent to Brigham Library, ETS, Princeton, NJ, 08541, USA, or via e-mail to [librarystaff@ets.org](mailto:librarystaff@ets.org).

“people of color” (acceptable). If the translators are unaware of such subtle aspects of the target language, the resulting guidelines may not be satisfactory.

### STEP 3: TRAIN THE PANELISTS

The first part of the meeting must be spent in training the panelists. Even though you distributed the manual to all of the participants before the meeting, it is necessary to discuss the following topics at the meeting to ensure that all of the panelists understand them.

**Purpose of Fairness Review:** Explain the purpose of fairness review to the panelists.

The primary purpose of fairness review is to identify invalid aspects of test items that might unfairly hinder people in various groups from demonstrating their relevant knowledge and skills. An additional purpose of fairness review is to identify invalid aspects of test items that might be perceived as unfair. The reviews should be based on clear and specific fairness review guidelines to make the reviews as objective and as comprehensive as possible.

**Fairness and Validity:** Explain the relationship between 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. Similarly, 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.

**Constructs and Variance:** Explain the meaning of the terms “construct” 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.<sup>5</sup> 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.<sup>6</sup> 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 measure. For example, if the language in the items is unnecessarily difficult, a test intended to measure quantitative reasoning may inadvertently also measure verbal ability. The verbal ability is a source of construct-irrelevant variance in a quantitative

---

<sup>5</sup> Variance is also a statistical term, but knowledge of statistics is not required to use this manual.

<sup>6</sup> No test meets the ideal, because there are always some elements of random chance in 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 scorers, thus causing some random score variance.

reasoning test. However, in a reading test, verbal ability would be a source of construct-relevant 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., students in a particular racial group, job applicants over forty years of age, 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 enhances the fairness and validity of assessments by removing identifiable construct-irrelevant sources of variance that may affect different groups in different ways.

**Principles:** Give the panelists an overview of the principles for fairness review of assessments before going into great detail on any single principle.

Taken together, the three principles cover the possible sources of construct-irrelevant variance: cognitive, affective, or physical.<sup>7</sup> The three sources of construct-irrelevant variance lead directly to the three principles for fairness review of 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 inches in 1.8 feet, knowledge of the relationship between feet and inches is a cause of construct-irrelevant variance. Test takers whose conversion skills are weak may answer the item incorrectly, even though they could have successfully multiplied 1.8 times 12.

If, however, the intended construct were conversion among units of length within the United States measurement system, then the need to convert feet to inches 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.

**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

---

<sup>7</sup> Cognitive sources of construct-irrelevant variance stem from differences in the knowledge bases of test takers. Affective sources of construct-irrelevant variance stem from differences in the emotional reactions of test takers. Physical sources of construct-irrelevant variance stem from differences in the abilities of test takers to see and hear test items and related material.

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 the 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 unnecessary 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.

**Skills Tests and Content Tests:** After you complete the discussion of the three principles, ask panelists to discuss the kinds of tests to which the principles will be applied. In the application of fairness review guidelines, it is important to distinguish between tests of general skills and abilities (skills tests) and tests of specific subject-matter knowledge (content 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, history, literature, nursing, or psychology. A content test may require material for valid measurement that would otherwise be out of compliance with the principles. A skills test would not require such material because the skill could be tested in many different contexts. For example, a detailed description of the gruesome effects of a car accident may be necessary in a content test for licensing emergency medical technicians, even though it would be unacceptable (in the United States) in a skills test of reading ability.

**Groups:** The guidelines apply to all test takers. Some groups, however, require special attention in the development and application of the guidelines because the members of such groups are more likely than others to have been objects of prejudice.

For example, the groups that received special attention in the development of the guidelines in the United States are defined by the following characteristics.

- Age
- Disability
- Ethnicity
- Gender
- National or regional origin
- Native language

- Race
- Religion
- Sexual orientation
- Socioeconomic status

Other groups may be of particular concern in tests made specifically for use in countries other than the United States. Ask panelists to identify the groups that are of particular concern in your country.

**Resolution of Disagreements:** Before the panelists begin to develop the guidelines, have the panelists discuss and agree on the criteria for resolving disagreements about the proposed guidelines. If panelists do not agree about a guideline, is acceptance by a simple majority of the group sufficient? Is acceptance by a larger proportion (e.g., two-thirds or three-quarters) of the group necessary? Should guidelines be limited only to those that all panelists accept?

**Need for Tolerance:** The final task before beginning work on the guidelines is to explain the need for the panelists to discuss sensitive topics.

It may be difficult for panelists to talk about such things as highly controversial topics, insulting stereotypes, and inappropriate labels for groups without inadvertently becoming offensive or being offended at times. The panelists should discuss that problem directly and reach an understanding of the mutual tolerance required to complete the delicate and important task ahead. In effect, panelists should excuse each other in advance for any inadvertently offensive statements that are made during the task of generating specific, locally appropriate fairness review guidelines.

#### **STEP 4: DEVELOP GUIDELINES WITH PANELISTS**

**Details of Each Principle:** The next set of tasks is to focus on the details of each one of the principles and to discuss questions that will lead to the generation of specific, locally appropriate guidelines.

The following principles for fairness in assessment apply to all tests made under ETS auspices. The operational implementation of each principle through the use of specific fairness review guidelines will vary from country to country, as appropriate for the culture(s) and customs of each country.

As a starting point, some of the guidelines in effect for ETS tests developed in the United States are described under each of the principles. Tell panelists that they may accept or reject any of those guidelines. Also, ask panelists to propose additional guidelines for discussion and possible adoption by the group.

## Principle 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. Principle 1 requires the avoidance of construct-irrelevant knowledge or skills.

Something that is construct irrelevant is not part of the knowledge, skills, abilities, or other characteristics that a test is supposed to measure. For example, the purpose of the question “How many nickels are in \$1.80?” is to measure the ability to divide decimal numbers. Given the purpose of the question, knowledge of United States coins is construct irrelevant. Test takers in many countries may be able to do the required division but may not know that a nickel is 0.05 of a United States dollar.

If construct-irrelevant knowledge is required to answer items in a test and the knowledge is not equally available to test takers, the validity and fairness of the test are reduced. The items should be revised to minimize the effects of the construct-irrelevant knowledge or skills.

A useful exercise to help panelists in the process of developing guidelines understand and identify construct-irrelevant knowledge or skills is to have the panelists review a sample of test items. Tell panelists as specifically as possible the knowledge or skills that each item is supposed to be measuring. Panelists should then discuss exactly what knowledge or skills are actually required to answer each item. Any discrepancies between what an item is supposed to be measuring and the knowledge or skills actually required to answer the item are construct-irrelevant knowledge or skills. For example, a problem in a mathematics test is supposed to measure quantitative skills. If the test takers have to write an explanation of how they arrived at their answers, however, verbal skills will be a source of construct-irrelevant variance.

**Translation:** Translation of test items without also accounting for cultural differences is a common source of construct-irrelevant knowledge. Translation alone may be insufficient for many test items, as shown by the example of an item that required knowledge of United States coins. The content of items must be adapted for the culture of the country in which the items will be used.

If tests have been translated from tests originally made for use in a different country, familiarity with the culture of that country may be a source of construct-irrelevant knowledge. If you are using translated tests, ask panelists to consider a guideline concerning the avoidance of construct-irrelevant topics that are specific to the country of origin of the test, such as brands of products, customs, entertainers, geography, government, history, holidays, institutions, laws, measurement systems, money, plants, politicians, political systems, sports, television shows, or wildlife. For example, an item could refer to the Fourth of July, which is an important holiday in the United States, but which may not be familiar to test takers in your country.

Because aspects of a language may vary in different countries in which the language is used, ask panelists to consider a guideline that would ensure that the version of the language used is appropriate for your country. If tests are given in English, for example, differences between American and British English in vocabulary and spelling may be a source of construct-irrelevant knowledge.

**Experience:** Discuss the following question with panelists and explain that the answer depends on the construct to be measured and on whether knowledge of the unfamiliar topic is required to answer the questions.

Is it acceptable to include materials that present concepts beyond the life experience of the intended population of test takers in a skills test? For example, is it acceptable to have a reading passage about snow in a reading test for students in a tropical country?

If one purpose of the reading test is to determine how well students can derive new information from printed text, then the use of unfamiliar topics is acceptable if all of the information necessary to comprehend the text is provided in the reading passage. It is not fair, however, to include passages about unfamiliar topics in a skills test if prior knowledge of the topic is required to answer the questions and the knowledge is not equally available to different groups of test takers.

**Words and Topics:** Following are words and topics that are likely to be sources of construct-irrelevant variance in the United States because knowledge of the words and topics is not spread evenly across various groups of people. For example, in the United States, men tend to know more about tools than do women. An item that requires knowledge of an uncommon tool is not fair unless the purpose of the item is to measure knowledge of tools.

Ask panelists to evaluate the list to determine whether any of the entries are likely to be sources of construct-irrelevant variance in your country. Even though some entries on the list may not apply in some countries, the list remains a useful way to focus the attention of panelists on potentially construct-irrelevant material.

Are the specialized words in Table 1 acceptable or unacceptable for your skills tests? What other fields of knowledge contain specialized words that should be avoided on skills tests in your country?

<b>Table 1. Examples of Common and Specialized Words</b>		
<b>Field of Knowledge</b>	<b>Acceptable Common Words</b>	<b>Unacceptable Specialized Words</b>
Digital technology	E-mail, computer, Internet	JPEG, MP3, server <sup>8</sup>
Farming	Field, harvest, plow	Thresher, tiller
Finance	Tax, salary, income	Arbitrage, hedge fund, preferred stock, venture capital
Law	Judge, jury	Subpoena, tort
Politics	President, vote	Earmark, filibuster, pork barrel
Science	Microscope, thermometer, degree	Lumen, vacuole
Tools and machinery	Hammer, engine	Torque, flange, chuck
Transportation	Train, car, sail	Catamaran, coupe

The following topics are likely sources of construct-irrelevant variance in the United States. Ask panelists to answer the questions listed for each topic to help them develop locally acceptable guidelines.

**Military topics.** Is knowledge of military topics such as conflicts, wars, battles, and military strategy acceptable in skills tests? Should knowledge about how military organizations function, knowledge of weapons, knowledge of the functions of parts of weapons, or knowledge of how weapons work be required on skills tests? Military topics are acceptable when they are construct relevant, as in some history tests.

**Regionalisms.** Should knowledge of words, phrases, and concepts more likely to be known by people in some regions of the country than in others be avoided in skills tests? If so, what are examples of such regionalisms that should be avoided?

**Religion.** Does the country have an official religion that all test takers are supposed to be familiar with? If not, it is probably best in any country 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.

<sup>8</sup> Terms in this field are likely to change status from specialized to common fairly rapidly.

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.

**Sports.** What are all test takers of skills tests expected to know about various sports? Will male and female test takers have the same level of knowledge?

Any particular knowledge or skill could be construct irrelevant in one test and yet be perfectly appropriate in another test. Specialized words associated with certain tools, for example, would be perfectly appropriate in a test used to license automobile mechanics but would be construct irrelevant in a test of reading skills. Check to be sure that the panelists are focusing on the avoidance of **construct-irrelevant** topics. Remind them that any topic is acceptable in a content test if it is important for valid measurement.

## **Principle 2. Avoid Affective Sources of Construct-Irrelevant Variance**

**Purpose:** The primary purpose of Principle 2 is to avoid emotional reactions caused by inappropriate, construct-irrelevant test content that may affect people in different groups in different ways.<sup>9</sup>

No group of test takers should have to face language or images that are unnecessarily contemptuous, derogatory, exclusionary, insulting, or the like. The contents of tests should not induce negative emotions that unnecessarily distract a test taker from the task of understanding a stimulus or responding to an item. In addition, construct-irrelevant aspects of a test should not make test takers feel alienated or uncomfortable. It is also 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 Principle:** How controversial, inflammatory, offensive, or upsetting does material have to be in order to violate Principle 2? Drawing a clear line with panelists 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 Principle 2. Consider the following factors in deciding whether material is in compliance with the principle.

**Age and sophistication of test takers.** Principle 2 is to be interpreted most strictly for young children. In general, the older and more sophisticated the test takers are, the more liberally the principle should be interpreted. For example, material that is unacceptable for high school students may be acceptable for college students.

---

<sup>9</sup> 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.

**Previous experience of test takers.** Tell panelists to consider the kinds of material that test takers are likely to have been exposed to extensively when they are deciding whether some test material is likely to offend or upset them. Brief prior 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.

**Importance for validity.** Explain to panelists that 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 Principle 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.

**Sensitive Topics:** Even though the particular topics will vary from country to country, it is likely in any country that some topics will be considered so sensitive that their use in tests should be avoided unless the topics are important for valid measurement. For example, in the United States, the topic of abortion is so controversial that it is best to avoid it in tests unless the topic is required for valid measurement, as might be the case in a test made for licensing nurses.

There are likely to be other topics that need not be avoided entirely in tests but that should be handled in a very careful manner. For example, in the United States, test developers should avoid dwelling on the horrible or shocking aspects of accidents or natural disasters, even though other aspects of those topics, such as the prevention of accidents, are acceptable in tests.

Following is a list of topics that have commonly been found to be in violation of Principle 2 for skills tests in the United States. For each entry, begin the discussion by asking panelists to answer the questions.

**Accidents, illnesses, or 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, but is it acceptable to dwell on gruesome, horrible, or shocking aspects of accidents, illnesses, or natural disasters in a skills test? Are other aspects of those topics acceptable? For example, is it acceptable to address the prevention of accidents, the causes of illness, or the occurrences of natural disasters?

**Advocacy.** Is it acceptable to use test content to advocate any particular cause or ideology, or should items and stimulus material be neutral and balanced whenever possible? Is it acceptable to take sides on any controversial issue when doing so is

not important for valid measurement? If so, which issues are acceptable? Which issues should be avoided?

**Death and dying.** Is it acceptable to focus on gruesome details associated with death and dying when doing so is not important for valid measurement? If not, what aspects of death and dying are acceptable on skills tests? Is a statement that someone died in a particular year or that a disease was responsible for a certain number of deaths acceptable?

**Evolution.** The topic of evolution has caused a great deal of controversy in the United States, and attitudes toward evolution vary greatly in different countries. What is the attitude toward evolution in your country? Should evolution be included in skills tests? How should evolution be treated in content tests?

**Group differences.** How should group differences be treated on skills tests? Is it acceptable to 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? Is it acceptable for any one group to be the standard of correctness against which all other groups are measured?

**Humor, irony, and satire.** It is acceptable to test understanding of humor, irony, and satire when it is important for valid measurement, as in some literature tests, but how should humor be treated on skills tests? Are all test takers likely to understand the joke, or might some be offended? How should irony and satire be treated? Could some people take such material literally and be offended by it? Is it important to avoid construct-irrelevant humor that is based on disparaging any groups of people, their strongly held beliefs, their concerns, or their weaknesses?

**Images.** An image or description of people and their interactions that is acceptable in some countries may be offensive to people in certain other countries. Is each of the following acceptable or unacceptable in your country?

- 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
- Images that do not conform to cultural norms, such as students behaving disrespectfully in the presence of an authority figure

Certain hand signals that are acceptable in some countries are inappropriate in other countries. Is each of the following acceptable or unacceptable in your country?

- 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)
- The thumbs-up sign

Are there other images, hand signals, or body language that should be avoided on tests in your country? For example, in some countries it is inappropriate to use the left hand for certain tasks.

**Imperialism.** In some countries, the topics of imperialism and colonialism are considered inappropriate in skills tests. Are those topics acceptable in skills tests in your country?

**Inadvertent references.** Materials used in tests come from many sources. Some of those sources may contain references to drugs, sex, racism, and other inappropriate topics. Be alert for such references and avoid them in tests unless they are construct relevant. Inadvertent references can be a problem because test developers may be unaware of the references. For example, in the United States, the date April 20 (Hitler's birthday) has become associated with racism. What inadvertent references to unsuitable topics may occur in materials selected for tests in your country? What symbols (e.g., swastikas) should be avoided?

**Luxuries.** Is it acceptable to depict 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, and the like?

**Personal questions.** Are there excessively personal questions regarding themselves, family members, or friends that should not be asked of test takers? Is asking about each of the following topics acceptable or unacceptable in your country?

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

**Religion.** What kinds of references to religion are allowed? For example, are the creation stories of various cultures allowed? Are references to religion, religious roles, institutions, or affiliations acceptable? If so, must all such references be to a particular religion or may different religions be mentioned? Is it acceptable to support or oppose religion in general or any specific religion? **In any case, it is not acceptable for an ETS test to disparage the members of any religion or their beliefs.**

**Sexual behavior.** In a skills test, how should sexual behavior be treated? What kinds of references to sexual behavior, if any, are allowed?

**Slavery.** Is slavery an acceptable topic? If so, what level of detail is acceptable? Would a discussion of the details of how slaves were packed in the holds of ships for transportation from Africa be acceptable in a skills test?

**Societal roles.** Are people in different groups found in a wide range of societal roles and contexts? If so, should language and images that suggest that all members of any single group are people in higher-status positions or lower-status positions be avoided? Would it be acceptable to have all the executives represented in a test be male and all the support staff be female? Is it acceptable to overrepresent members of any group in examples of inappropriate, foolish, unethical, or criminal behavior?

**Stereotypes.** ETS test developers are told to avoid stereotypes (both negative and positive) in language and images unless important for valid measurement. Are stereotypes acceptable in your country? Is it acceptable to use phrases that encapsulate stereotypes, such as the English phrases “women’s work,” or “man-sized job?” Is it acceptable to use words such as “surprisingly” or “amazingly” when the surprise or amazement is caused by a person’s behavior contrary to a stereotype? For example, would a sentence such as “Surprisingly, a girl won first prize in the science fair” be acceptable?

The terms “stereotypical” and “traditional” overlap in meaning but are not synonymous. A traditional activity (such as a woman cooking) does not necessarily constitute a stereotype as long as the test as a whole does not depict members of a group engaged exclusively in traditional activities. In the United States, if some group members are shown in traditional roles, other group members should be shown in nontraditional roles. What balance of traditional and non-traditional roles is acceptable in a skills test in your country?

**Substance abuse.** Is a focus on the details of substance abuse acceptable? Are alcohol, tobacco, and prescription (as well as illegal) drugs included among abused substances?

**Suicide or other self-destructive behavior.** Is it acceptable to mention that a person committed suicide? Is it acceptable to focus on various means of suicide or to glorify suicidal behavior or other self-destructive behavior?

**Topics best avoided.** Some topics are considered to be so problematic in the United States that they have been avoided in skills tests made for use in that country. Which, if any, are unacceptable for skills tests in your country?

- 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

What other topics are so problematic in your country that they should be avoided unless they are important for valid measurement in content tests?

**Unstated assumptions.** Is it acceptable to use material based on underlying assumptions that are false or that would be inappropriate if the assumptions had been stated? For example, would material that assumes all children live in houses with backyards be acceptable?

In some countries, 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. Is it acceptable to use an undefined “we” in your country?

**Violence and suffering.** In skills tests, is it acceptable to focus on violent actions, on the detailed effects of violence, or on suffering? Is it acceptable to discuss the food chain, even though animals are depicted eating other animals?

Ask panelists to add topics of concern in your country that do not appear among the topics listed above. For example, in some countries, discussion of anything negative about the royal family would be considered inappropriate.

**Use Appropriate Terminology for Groups:** If group identification is necessary, it is usually most appropriate to use the terminology that group members prefer. Do not use derogatory names for groups, even if some group members use them to refer to other members of their group. Inflammatory terms should be avoided.

Ask panelists to decide which groups should be mentioned in the test and determine how to refer to each group appropriately. As an example of one useful way to address the issue, the terminology that is acceptable in the United States is described below.

In the United States, it is preferable to 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. Is this true in your country?

Discussions of appropriate terminology for various population groups follow. Ask panelists to specify the appropriate terminology for your country.

**Asian people.** ETS test developers are told to use specific terminology such as “Chinese” or “Japanese” and not to use the word “Oriental” to describe people unless quoting historical or literary material or using the name of an organization.

How should Asian people be referred to in your country?

**Bisexual, gay, lesbian, or transgendered people.** In the United States, people should be identified by sexual orientation only when it is important to do so for valid measurement. Is that true in your country as well? The words “bisexual,” “gay,” “lesbian,” and “transgendered” are all acceptable. ETS test developers are told to avoid using the term “homosexual” outside of a scientific, literary, or historical context. They are told to avoid 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. They use the phrase “sexual orientation” rather than “sexual preference.”

How should bisexual, gay, lesbian, or transgendered people be referred to in your country?

**Black people.** In the United States, “Black” and “African American” (for American people) are both acceptable. “Negro” and “colored” are not acceptable except in historical material or the official names of organizations. Because “Black” is used as a group identifier, ETS test developers are told to avoid the use of “black” as a negative adjective, as in “black magic,” “black day,” or “black-hearted.”

How should Black people be referred to in your country?

**People with disabilities.** Following are the guidelines used in the United States regarding people with disabilities. The guidelines may or may not be appropriate in your country. Ask panelists to review the guidelines and decide which should be retained and which should be deleted. Then ask panelists to decide if any guidelines should be added.

To avoid giving the impression that people are defined by their disabilities, ETS test developers are told to focus on the person rather than the disability in the first reference to someone with a disability. In the United States, the preferred usage is to put the person first and the disabling condition after the noun, as in “a person who is blind.” Is that appropriate in your country?

With respect to people with disabilities, ETS test developers are told to avoid the following.

- 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.”
- Euphemistic or patronizing terms such as “special” or “physically challenged”
- Achieving success “in spite of” a disability or “overcoming” a disability

- 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.
- Implying that someone with a disability is sick (unless that is the case) by the use of 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, “client” is the appropriate term.)

Table 2 distinguishes among appropriate and inappropriate English terms related to disabilities, as used in the United States. How should those terms be handled in your country? If the test is not in English, consider what the equivalent terms would be in the language of the test.

<b>Table 2. Use of Terms Related to Disabilities</b>	
Abnormal, normal	Unacceptable in most cases. The terms “normal” and “abnormal” are not appropriate for referring to people except in biological or medical contexts.
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.

Should any terms be added to, or deleted from, the table for use in your country?  
Should any of the terms in the table be modified?

**Hispanic people.** ETS test developers are told that “Hispanic” and “Latino” (for men) or “Latina” (for women) are acceptable but that it is preferable to use a specific group name such as “Cuban,” “Dominican,” or “Mexican,” when possible.

How should Hispanic people be referred to in your country?

**Indigenous people.** ETS test developers are told that whenever possible, it is best to refer to indigenous 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 Diné are still known to many other peoples as the Navajo.” Many indigenous people prefer the words “nation” or “people” to “tribe.”

How should indigenous people be referred to in your country?

**Members of minority groups.** In the United States, members of minority groups are becoming the majority in many locations. Minority groups in the United States are the majority in many other countries. Therefore, although the terms are still

acceptable, ETS test developers are told to try to reduce the use of “minority” and “majority” to refer to groups of people. They are told to avoid the terms “disadvantaged minorities” and “underserved minorities” because the terms are vague and loaded with unstated assumptions. They are told to use the names of the specific groups the terms are intended to include, instead.

How should members of minority groups be referred to in your country?

**Multiracial people.** The terms “biracial” and “multiracial,” as appropriate, are acceptable in the United States for people who identify themselves as belonging to more than one race. In the United States, “people of color” is acceptable for biracial and multiracial people and for people who are African, Asian, Hispanic, or Native American. “Colored people” is not acceptable except in historical or literary material or in the name of an organization.

How should multiracial people be referred to in your country?

**Older people.** In the United States, it is considered 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.” ETS test developers are told to avoid using “elderly” as a noun and to 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.

How should older people be referred to in your country?

**People below the poverty level.** The term “poor” is vague and ambiguous. It is preferable to use specific income ranges. ETS test developers are told not to use “poor” as a noun except when quoting literary or historical material and not to 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.

How should people with insufficient incomes be referred to in your country?

**White people.** In the United States, the terms “White” and “Caucasian” are both acceptable, but “White” is becoming the preferred term.

How should White people be referred to in your country?

**Women and men.** In the United States, women and men must be referred to in parallel terms. ETS test developers are told the following.

- 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 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.
- In English, 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. Role labels such as “scientist” or “executive” include both men and women.
- Expressions such as “the soldiers and their wives” that assume only men fill certain roles should not be used unless such is the case in some particular instance.
- Generic role words should not be coupled with gender-specific pronouns or actions unless a particular person is being referenced. For example, terminology that assumes all kindergarten teachers or food shoppers are women or that all college professors or car shoppers are men should not be used.
- Objects should not be referred to using gender-specific pronouns except in historical or literary material.
- Using “he” or “man” to refer to all people is not acceptable. The use of generic “he” or “man” is not acceptable 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.<sup>10</sup>
- The constructions “he/she” and “(s)he” should be avoided.

How should men and women be referred to in your country?

Table 3 gives examples of unacceptable usages of “man” and “he” in the United States along with acceptable alternatives. Are the same generic male references unacceptable in your country? Are the same alternatives acceptable? What changes should be made to Table 3 for use in your country?

---

<sup>10</sup> 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.

<b>Table 3. Alternatives to Use of Generic Male References</b>	
<b>Unacceptable</b>	<b>Acceptable</b>
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, mailman, 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
Manpower	Workers, personnel, labor, workforce

**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, ETS test developers strive to obtain at least the level of diversity described below in tests that mention or show people.

The diversity in tests made for your country should reflect, to the extent possible, the diversity of the test-taking population. Ask panelists to determine the groups in the test-taking population to be included in a test made for your country. The guidelines for the United States are shown below as potentially helpful models for the panelists to consider.

**Racial and ethnic balance.** For skills tests, ETS test developers 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, ETS test developers strive to

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. They also include indigenous groups from the United States or from other countries. Items that include other groups that could be considered minorities, such as Americans of Middle Eastern origin or Middle Eastern people, may be counted among the items that represent diversity.

If there is insufficient context in an item to indicate group membership in other ways, representation may be accomplished by using the names of reasonably well-known real people in various groups or by using generic names commonly associated with various groups. ETS test developers are told not to add unnecessarily to the linguistic loading of the item by using names that are inordinately difficult for test takers to decode.

Do any of these guidelines apply in your country? What other guidelines should apply? What proportion of the items in skills tests should represent people from what are commonly considered to be minority groups in your country?

In some content tests, such as history 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, ETS test developers are told to 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.

ETS test developers try to achieve at least a rough parity in status of the people depicted in different racial and ethnic groups. They do not, for example, depict all managers as White and all workers as Black or Hispanic. Do any of these guidelines apply in your country? What other guidelines should apply?

ETS test developers are told to occasionally represent people with disabilities in tests that include people, but to be careful not to reinforce stereotypes when doing so. 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.

Would any of the guidelines or similar guidelines apply in your country? Should any guidelines be added?

**Gender balance.** In the United States, 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.<sup>11</sup> 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.

Do any of these guidelines apply in your country? What other guidelines should apply?

Ask panelists to establish guidelines for the percentages of items that should represent various groups in the test-taking population. The percentages may be limited to only those items that mention people. For example, if half of the test-taking population is female, the guidelines for skills tests may indicate that about half of the items that mention people should refer to women or girls.

### **Principle 3. Avoid Physical Sources of Construct-Irrelevant Variance**

**Purpose:** The purpose of Principle 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.

**Types of Barriers:** Explain to the panelists that there are three types of physical barriers.

**Essential Barriers.** Some 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 Barriers.** 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 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

---

<sup>11</sup> Some representation of both genders is required, however.

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 Barriers.** Some physical barriers, however, are simply not necessary. They are not essential to measure the construct, nor are they even helpful. 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 Principle 3 is on the avoidance of unnecessary physical barriers in items and stimuli.

**Developing Guidelines:** 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.

These guidelines are likely to apply in all countries, but ask panelists to review the guidelines to make sure that they are all applicable. The panelists should modify the guidelines as necessary. For example, if your country does not use the Roman alphabet, the guideline about letters that look alike would have to be changed to reflect the writing system in use. The need to avoid characters that look alike as labels in the same item is likely to be universal, however.

ETS test developers are told to avoid the following barriers, or others like them, if they are neither essential nor helpful for measuring the intended construct.

- Construct-irrelevant charts, maps, graphs, and other visual stimuli
- Construct-irrelevant drawings of three-dimensional solids, such as adding a meaningless third dimension to the bars in a bar graph
- Construct-irrelevant measurement of spatial skills (visualizing how objects or parts of objects relate to each other in space)
- Decorative rather than informative illustrations
- Visual stimuli 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<sup>12</sup>
- Numbers 1–10 and letters A–J used as labels for different things in the same item or stimulus (because the same symbols are used for those numbers and letters in Braille)
- Special symbols or non-English alphabets (unless that is standard notation in the tested subject, such as  $\Sigma$  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)

Would any of the guidelines apply in your country? Which guidelines would apply with some revision, e.g., changes to letters that sound alike? Which other guidelines should apply?

Test developers in all countries need to 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. To the extent possible, test developers should reduce the need to scroll down the computer screen to access parts of stimulus material, unless the ability to scroll is construct relevant.

### **Additional Requirements for Tests for School Children**

**Rationale:** In the United States, tests designed for school children in kindergarten through grade 12 (K–12) are usually subject to additional guidelines for fairness review. Various constituent groups may have very strong beliefs about acceptable test content for their children, and those beliefs are reflected in the fairness review guidelines for K–12 tests.

Is the same true in your country? If so, ask panelists to review the K–12 guidelines and determine which, if any, should apply to K–12 tests in your country.

**Additional Requirements:** The following requirements for United States K–12 tests are all extensions of Principle 2 (Avoid affective sources of construct-irrelevant variance). Any redundancy with the discussion of Principle 2 is intended to stress the additional care to be employed in tests for young students.

In the United States, ETS test developers are told to avoid items or stimuli about each of the following in K–12 tests, unless important for valid measurement. Should the same topics be avoided in your country?

---

<sup>12</sup> This does not apply to the traditional option labels (A–E) for multiple-choice items, even though B and D sound alike.

- Serious illnesses (e.g., cancer, HIV, AIDS, herpes, tuberculosis, smallpox, anthrax)
- Animals or people who are killed or are dying
- Natural disasters, such as earthquakes, hurricanes, 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
- Divorce, loss of jobs, layoffs, and other family situations that students may find upsetting
- Depictions or suggestions of interpersonal violence or disharmony, including playground arguments, fights among students, bullying, cliques, and social ostracism
- Dissension among family members or between students and teachers
- Graphic depictions of violence in the animal kingdom
- 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)
- Drinking alcohol, smoking or chewing tobacco, and gambling
- Drug use, including the use of prescription drugs
- Birthday celebrations, Christmas, Halloween, and Valentine's Day (because not all children celebrate them)
- Dancing, including school dances (such as proms), particular kinds of music (e.g., rap, rock and roll)
- Going to the movies
- References to a deity, including expressions like "Thank God," and euphemisms for such references (for example, "gee whiz")
- Stories about mythological gods or creation stories
- Extrasensory perception, UFOs, the occult, or the supernatural
- Texts that are preachy or moralistic, as they may offend populations that do not hold the values espoused
- Controversial topics such as 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
- Evolution, with associated topics of natural selection, fossils, geologic ages (e.g., millions of years ago), dinosaurs, and similarities between people and primates
- Particular personal or political values in discussions of controversial topics such as protection of the environment, deforestation, or labor unions
- 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.)
- Materials that model or reinforce inappropriate student behaviors such as students playing tricks on teachers or adults, lying, stealing, or running away from home, or even considering those behaviors
- Students going without sleep, failing to attend school or do homework, or eating large quantities of junk foods
- Students violating good safety practices (e.g., students keeping dangerous animals, entering homes of unknown adults), even if everything turns out well

- Suggestions of sexual activity and words or phrases that carry a sexual connotation
- Children coping with adult decisions or situations (e.g., supporting the emotional needs of a parent)
- Expressing or implying cynicism about charity, honesty, or similar values

Should any of these topics, or any additional topics, be avoided in K-12 tests in your country?

## **STEP 5: ESTABLISH PROCEDURES**

**Essential Aspects of Fairness Review:** ETS has established procedures that it applies to all fairness reviews. Those procedures are listed below. Ask the panelists to review the ETS procedures and establish procedures to be implemented in your country.

- Item writers, reviewers, and editors should be trained to follow the fairness review guidelines.
- All items should be reviewed for fairness by trained fairness reviewers before the items are used in tests.
- To the extent possible, the fairness reviewers should have no stake in the test being reviewed. Item writers cannot serve as reviewers of items they have written themselves. Test developers who submit items for review should not be able to select the particular fairness reviewers who will review their items.
- The fairness reviewer should 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 any audiotapes (or scripts) and visual materials, in addition to the items.
- The fairness review should be documented.
- Items or materials that have been challenged by a fairness reviewer should not be used until the challenge has been resolved. The resolution should be documented.
- Material that is very expensive to change at later stages (e.g., videotapes, extended reading passages) should receive a fairness review before any substantive work is done. The review of expensive stimulus materials is strongly recommended as a way to reduce the risk of expending resources on materials that later may be found to be out of compliance with the guidelines. The fairness review of items based on the stimulus remains mandatory.

Ask panelists to develop adjudication procedures for use in case there are disputes about the results of a fairness review. For example, the panelists may decide to establish a fairness committee to resolve any disputes that occur between item writers and fairness reviewers.

## **STEP 6: COMPLETE FAIRNESS REVIEW GUIDELINES**

The outcome of the meeting should be detailed notes about the panelists' decisions. The notes must be transformed into a set of clear and detailed guidelines that will help test

developers create items and tests that are fair for all of the intended test takers in your country.

As noted earlier, ETS believes that obtaining the opinions of many diverse people is an appropriate way to develop fairness review guidelines. Therefore, widespread review of the draft document is recommended. The designated author(s) should write a draft of the fairness review guidelines that will be reviewed by all of the panelists first. After all of the panelists are satisfied that the document accurately reflects the decisions that they made, send the document to the reviewers identified at the beginning of the process. The more reviewers you have and the more diverse the reviewers are, the less likely you are to be surprised by criticism of the finished guidelines.

It is likely that the document will require revision based on the reviews before it becomes operational. It is likely that two or three rounds of revision and review will be necessary before the document is completed and ready for review and approval by the previously identified decision maker(s).

The result should be fairness review guidelines based on the principles, customized to reflect the specific fairness concerns of your country, and detailed enough to guide the test-creation process.<sup>13</sup>

Test developers who are interested in *ETS Guidelines for Fairness Review of Assessments* (2009) may download the document at no cost from [www.ets.org](http://www.ets.org). Even though much of the content of the *Guidelines* deals primarily with fairness issues in the United States, the document is a useful model of how the principles for fairness in assessment have been supplemented with explanations and examples to serve as specific fairness review guidelines.

## **STEP 7: TRAIN TEST DEVELOPERS, REVIEWERS, AND EDITORS**

People who will be involved in the test-development process in your country must be trained to use your fairness review guidelines. In the experience of ETS, simply reading the guidelines document is not sufficient. The people who will write, review, or edit test items should be given the opportunity to apply the guidelines to samples of items, to discuss the results with a group of their colleagues, and to attempt to resolve any differences in their interpretations of the guidelines.

Most of the items used in the training sessions should be carefully selected to present subtle fairness problems. Obviously unfair items are helpful only in the earliest stages of training. Items that cause disagreements among the trainees are the most useful training items. The training should discourage both excessively zealous and excessively lax interpretations of your guidelines.

---

<sup>13</sup> ETS requests that a copy of your guidelines, or a URL from which the document may be downloaded, be sent to Brigham Library, ETS, Princeton, NJ, 08541, USA, or via e-mail to [librarystaff@ets.org](mailto:librarystaff@ets.org).

## **CONCLUSION**

Keep in mind that it is impossible to develop rules and examples for fairness review that will cover every situation. Experience will surely lead to revisions of your guidelines. Furthermore, what is considered fair changes over time, so some aspects of your guidelines will eventually become obsolete. It is a good practice to schedule a review of your fairness review guidelines every five years or so.

The development of clear and specific fairness review guidelines appropriate for use in a particular country is a complex and time-consuming task. The people who will use the guidelines must be trained, and the guidelines must then be applied to all test items and to all test-related materials. The benefit of fairer and more valid tests for all of the people who take them, however, is well worth the time and effort involved.

## SOME USEFUL REFERENCES ON FAIRNESS IN ASSESSMENT

- 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. 2001. *Publication manual of the American Psychological Association*. Washington, DC: APA.
- 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. L. Brennan. Washington, DC: American Council on Education.
- 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.
- Educational Testing Service. 2009. *ETS guidelines for fairness review of assessments*. Princeton, NJ: ETS.
- Holland, P., and H. Wainer, eds. 1993. *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- 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.
- 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, MN: University of Minnesota, National Center on Educational Outcomes.
- Zieky, M. J. 2006. Fairness reviews in assessment. In *Handbook of test development*, eds. S. Downing and T. Haladyna. Mahwah, NJ: Lawrence Erlbaum.



*Listening. Learning. Leading.®*

*[www.ets.org](http://www.ets.org)*