Measuring the Power of Learning.®

ETS International Principles for the Fairness of Assessments

A Manual for Developing Locally Appropriate Fairness Guidelines for Various Countries

By Educational Testing Service

ETS International Principles for the Fairness of Assessments

Educational Testing Service

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 Fair Tests and Communications* (ETS, 2015). The document may be downloaded at no cost from www.ets.org.

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. The use of guidelines for the fairness of tests 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 the Fairness of Assessments* supports this mission by helping to ensure that tests created 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 the fairness of tests 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 that will enable them to build assessments that are fair for the intended test takers within the country.

I am pleased to issue the 2016 version of the *ETS International Principles for the Fairness 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 the Fairness of Assessments* is to help you ensure that tests made to be used in a specific country will be fair for test takers in that country. This manual is intended to help you avoid the inclusion of unfair content in tests as the tests are developed, and eliminate any inadvertently included unfair content as the tests are reviewed. The focus of the manual is on fairness with respect to test content, not on psychometric or statistical measures of fairness, and not on issues related to the fairness of other aspects of the testing process or the use of test scores.

Definition of fairness. For test developers, the most useful definition of fairness in assessment is based on validity. Validity is the most important indicator of test quality. Messick (1989, p. 13) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores" (emphasis in the original).

According to the S*tandards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 49), "fairness is a fundamental validity issue." Fairness in the context of assessment can usefully be defined as the extent to which inferences and actions based on test scores are valid for diverse groups of test takers.

Valid test content is relevant to the intended purpose of the test. Relevant (valid) test content is necessarily fair. Irrelevant (invalid) test content may or may not be fair. If irrelevant content affects all test takers to about the same extent, validity is diminished. If irrelevant content affects some group of test takers (e.g., women) more than some other group of test takers (e.g., men), then fairness is diminished as well as validity.

Rationale. ETS recognizes that what is considered fair test content will vary from country to country. ETS is not attempting to impose its specific fairness guidelines, which were designed for use primarily in the United States, on other countries. For example, tests designed for use in Qatar are likely

to require a different set of guidelines regarding potentially offensive topics than are required for tests designed for use in the United States.

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. Although the focus of this manual is on tests, the concepts discussed also apply to related documents such as test descriptions, practice materials, administrator's manuals, and essay scoring guides.

Regardless of the local guidelines that are set, no test should contain material that expresses or incites hatred or contempt for people on the basis of age, atypical appearance, citizenship status, disability, ethnicity, gender (including gender identity or gender representation), national or regional origin, native language, race, religion, sexual orientation, or socioeconomic status.

Overview. In this manual we describe the universal principles for fairness and then describe how to generate locally appropriate fairness guidelines based on those principles. We next discuss how to establish procedures for use of the guidelines, how to train users of the guidelines, and how to apply, monitor and revise the guidelines.

We include samples of fairness guidelines that ETS developed primarily for the United States. The specific guidelines are not necessarily recommended for use in countries other than the United States. Those guidelines are intended only to stimulate discussion about locally appropriate guidelines. We conclude with a brief description of additional actions that should be taken to help make tests as fair as possible, and a list of books and articles relevant to various aspects of fairness in assessment.

UNIVERSAL PRINCIPLES

Though particular guidelines will vary from country to country, there are general principles for fairness that appear to be universal.

Measure the important aspects of the relevant content. A test that does not measure the important aspects of the intended content cannot be valid. Because of the close link between validity and fairness, an invalid test is not likely to be fair. Therefore, any material that is important for valid measurement may be acceptable for inclusion in a test, even if it would otherwise be out of compliance with the guidelines. Some offensive or upsetting material may be important in certain content areas. A history test, for example, may appropriately include material that would otherwise be out of compliance with the guidelines to illustrate certain attitudes commonly held in the past. Professional judgment is required to evaluate the importance of the material for valid measurement against the extent to which the material may act as an unfair barrier to the performance of some test takers.

Avoid irrelevant cognitive barriers to the performance of test takers. Unfair barriers may occur when knowledge or skill not related to the purpose of the test is required to answer an item correctly. For example, if an item that is supposed to measure multiplication skills asks for the number of meters in 1.8 kilometers, knowledge of the relationship between meters and kilometers is irrelevant to the intended focus of measurement. Test takers whose conversion skills are weak may answer the item incorrectly, even though they could have successfully multiplied 1.8 times 1,000. If, however, the intended purpose were to measure conversion among units within the metric system, then the need to convert kilometers to meters would be relevant and, therefore, fair.

Avoid irrelevant emotional barriers to the performance of test takers. Unfair barriers may occur if unnecessary language or images cause strong emotions that may interfere with the ability to respond to an item correctly. For example, offensive content may make it difficult for test takers to concentrate on the meaning of a reading passage or the answer to a test item, thus serving as an irrelevant barrier to performance. 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.

Avoid irrelevant physical barriers to the performance of test takers. Unfair barriers may 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 and respond to them. 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 knowledge and skills that are supposed to be tested by the item based on the diagram.

DEVELOP GUIDELINES

Start early. Ideally, the development of specific fairness guidelines based on the universal principles should take place before the test development process begins. The people who write and review test items, and those who assemble and review tests should all be familiar with the fairness 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 for compliance with the guidelines before the items are administered to test takers.

It will probably take several months to complete the process of developing locally appropriate guidelines. We recommend pooling the opinions of diverse people to help you develop the guidelines. You will need time to discuss what the guidelines should be with those people, and time to write the resulting fairness guidelines. Then additional time will be required to have the draft guidelines reviewed, revised, and accepted. The people who will use the guidelines have to be trained to use them. Finally, the guidelines should be monitored and reviewed periodically and updated as needed.

Obtain help. While it is possible for a well-informed individual to write fairness guidelines, we believe that the task of augmenting the general principles to form specific guidelines is best accomplished by a diverse group of people who are very familiar with your country and who are also familiar with the

intended population of test takers. Therefore, in this manual we assume that you are working with such colleagues. For tests made for use in your country's schools, include teachers among the people helping you because knowledge of curricula and instructional practices is important for evaluating the fairness of such tests.

It is helpful to include people who represent the important subgroups of the country's population to the extent possible. For example, 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 to the extent possible, and so forth.

Before you begin to work on the guidelines with your colleagues, explain the need to feel free to discuss sensitive topics. It may be difficult to talk about such things as highly controversial topics, insulting stereotypes, and inappropriate labels for groups without inadvertently becoming offensive at times. Discuss that problem directly and reach an understanding of the mutual tolerance required to complete the delicate and important task ahead.

Sample guidelines. 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 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.

The sample guidelines may or may not be appropriate for a test made specifically for a country other than the United States. In developing local guidelines, you may accept, modify, or reject any of those sample guidelines. The sample United States guidelines are not likely to cover all of the important fairness issues in the country for which the test is being made. Additional guidelines are likely to be necessary to cover issues specific to the country. We raise questions about the sample guidelines for your consideration.

Groups of primary concern. The guidelines apply to all test takers. Some groups, however, require special attention in the development and application of fairness guidelines because the members of such groups are more likely than others to be discriminated against.

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

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

What characteristics define the groups that should receive special attention in the development of your guidelines? For example, in some countries the type of school a test taker attended could be a relevant factor.

Irrelevant Cognitive Barriers

Language. Language that is more difficult than is necessary for valid measurement is a common source of irrelevant cognitive barriers to performance. Use the most accessible level of language that is consistent with valid measurement. While the use of accessible language is particularly important for test takers who have limited skills in the language of the test, the use of such language is beneficial for all test takers when linguistic competence is not part of what is being measured.

Avoid requiring knowledge of excessively specialized vocabulary unless such vocabulary is being assessed on purpose. Do not require knowledge of words, phrases, and concepts more likely to be known by people in some regions of the country than in others (e.g., dialects and certain idioms), unless it is important for valid measurement. What is considered excessively specialized or regional requires judgment. Take into account the maturity and educational level of the test takers in deciding which words are too specialized.

Difficult words and language structures may, of course, be used if they are important for validity. For example, difficult words may be appropriate if the purpose of the test is to measure depth of general vocabulary or specialized terminology within a subject-matter area. It may be appropriate to use a difficult word if the word is defined in the test or its meaning is made clear by context. Complicated language structures may be appropriate if the purpose of the test is to measure the ability to read challenging material.

What level of vocabulary and syntax is acceptable for the tests you are developing? How would you describe "accessible language" for item writers to use? What aspects of language should item writers avoid unless language is the intended focus of measurement?

Topics. It is necessary to avoid requiring irrelevant, specialized knowledge to answer an item correctly. For example knowing the number of players on a rugby team would be relevant on a licensing test for physical education teachers, but not on a mathematics test.

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

ETS identified certain subjects as likely sources of irrelevant specialized knowledge in the United States. For example, irrelevant knowledge of sports, the military, and tools tended to make items more difficult for women than for men at the same level of knowledge and skill in the tested subject. The

sources of irrelevant specialized knowledge in tests made specifically for use in countries other than the United States are likely to be different.

What aspects of specialized knowledge that are not the point of measurement are likely to be irrelevant cognitive barriers in your country?

Translation. Translation of test items without also accounting for cultural differences is a common source of barriers to performance related to measurement of irrelevant knowledge. Translation alone may be insufficient for many test items. The content of items must be adapted for the culture of the country in which the items will be used. For example, an item in a test originally made for use in the United States could refer to the Fourth of July, which is an important holiday there, but which may not be familiar to test takers in other countries. If you are using translated tests, consider a guideline concerning the avoidance of irrelevant topics that are specific to the country of origin of the test.

Translation issues may exist even if the same language is used in various countries. For example, if tests are given in English, differences between American and British English in vocabulary and spelling may be a source of irrelevant knowledge.

ETS identified the following topics as potentially requiring irrelevant knowledge when tests originally made for use in the United States are used in other countries.

- brands of products, names of corporations,
- celebrities, entertainment, sports, and television shows,
- culture and customs,
- geography,
- government, politics, and politicians,
- history,
- holidays,
- institutions,
- laws,
- measurement systems, and units of money,

- plants and wildlife peculiar to the United States.

Which topics would be of concern in translated tests used in your country? What additional topics would be of concern?

Contexts. In tests that measure skills rather than content knowledge (e.g., reading comprehension), stimuli, such as reading comprehension passages, still have to be about some content. Similarly, applications of mathematics generally require some real-world setting. The contents of reading passages and the settings of mathematics problems have raised fairness issues. It is not appropriate to assume that all test takers have had the same experiences. Is it fair to have a reading passage about snow when students in tropical countries may have never experienced it? What contexts are fair to include in tests?

The answer depends on what test takers in a particular grade are expected to know about the context, and on the extent to which the information necessary to understand the context is available in the stimulus material. Generally, school-based experiences are more commonly shared among students in a particular grade than are their home or community-based experiences. In any case, a very important purpose for reading is to learn new things. It could severely diminish validity to limit the contents of reading passages to content already known by test takers.

If reading comprehension is to be measured rather than knowledge of the subject matter from which the passage is excerpted, then the information required to answer the items correctly should either be common knowledge among the intended test takers or be available in the passage. Similarly, for mathematics problems, the contexts should be common knowledge among the intended test takers, or the necessary information should be available in the problem. The teachers of the relevant grades are a very helpful source of information about what is considered common knowledge at those grades.

For test takers with disabilities, there is an additional requirement that direct, personal experience unavailable to the disabled test takers not be required to understand the context. For example, a test taker who is unable to participate in a footrace can still understand a problem set in the context of a footrace. On the other hand, a passage about the emotional impact of colors may be inappropriate for test takers

who have never been able to experience colors. Therefore, it is best to avoid irrelevant contexts that require direct, personal experience to be understood, if those experiences are not available to people with certain disabilities.

What contexts are likely to be appropriate for the tests you are developing? Are there contexts that should be avoided?

Religion. Religion is a common source of irrelevant knowledge. For tests made primarily for use in the United States, ETS has told test developers to use only the information about religion that is important for valid measurement.

How should religion be treated in tests in your country? Is there some knowledge about religion that all test takers are assumed to have, or should religion be avoided unless it is the focus of measurement?

Guidelines Regarding Irrelevant Emotional Barriers

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, irrelevant aspects of a test should not make test takers feel alienated or uncomfortable. It is also important to avoid irrelevant content that is commonly believed to be unfair, even if it is not certain that test takers' performance is actually affected by such content.

Advocacy. Items and stimulus material should be neutral and balanced whenever possible. Do not use test content to advocate any particular cause or ideology nor take sides on any controversial issue unless doing so is important for valid measurement. For example, in the United States gun control has become highly divisive. If material in a test argues either for or against gun control, some test takers will be angered by the material and their responses to the test could be adversely affected.

What topics are so divisive in your country that advocacy of one side or the other should be avoided in tests unless required for valid measurement?

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.

What topics are so sensitive in your country that it is best to avoid them in tests unless they are required for valid measurement? For example, in some countries criticism of the royal family must be avoided.

There are likely to be other topics that need not be avoided 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.

Some topics that have commonly been found to be sources of emotional barriers in the United States are contraception, euthanasia, evolution, sexual issues, extreme violence, and slavery. Any list of troublesome topics can be only illustrative. Current events, such as a highly publicized terrorist attack or a destructive natural disaster, can cause new topics to become distressing at any time. It is a good practice to obtain a fairness review of any potentially problematic material before time is spent developing it for use on a test.

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

In your country, what topics must be handled with care because they are likely to present emotional barriers to the performance of test takers?

Stereotypes. Avoid stereotypes (both negative and positive) in language and images unless important for valid measurement. Do not imply that all members of a group share the same characteristics, unless the group was assembled on the basis of those characteristics.

What stereotypes should be avoided in tests in your country?

Appropriate terminology for groups. If ethnic, gender, or racial group identification is necessary, it is generally most appropriate to use the terminology that group members prefer. Avoid derogatory terms. In authentic historical and literary material, some violations of the guidelines may be present. Even in such material, however, offensive terms should be avoided unless they are important for valid measurement.

Which groups may be of concern regarding appropriate terminology in your country? For each group, describe the terminology that is appropriate to identify the group in your country.

Representation of diversity. If a test mentions or shows people, test takers should not be made to feel alienated from the test because no members of their group are included. Therefore, the ideal test would include members of the various relevant groups in the test-taking population. While it is not feasible to include members of every relevant group in a test, strive to represent diversity in tests that mention or show people. The diversity reflected in tests made for a specific country other than the United States should be appropriate for the country for which the test is designed.

Which groups should be represented in the tests in your country? Approximately what proportion of items that mention people should be allocated to representing diverse groups?

Additional Requirements for Tests for School Children

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

Unless the topics are important for validity, avoid a discussions of topics that may be particularly emotionally charged for K–12 students. For example test developers in the United States are told to avoid

material about the death or serious illness of parents or siblings, as well as family problems such as divorce or loss of a breadwinner's job. The test developers are told to avoid topics that may be offensive to certain groups such as use of alcohol or gambling. They are also told to avoid materials that model or reinforce inappropriate student behaviors such as lying, stealing, or cheating.

Are special guidelines needed for K-12 tests in your country? If so, which topics should be avoided unless they are required for valid measurement?

Guidelines Regarding Irrelevant Physical Barriers

Tests and related materials should be created in formats that are accessible to individuals with disabilities, to the extent possible. Even if that is done, however, some physical barriers may remain and some test takers with disabilities may still need accommodations.

Irrelevant physical barriers occur if aspects of tests not important for validity interfere with the test takers' ability to attend to, see, hear, or otherwise sense the items or stimuli and/or to enter a response to the item. For example, test takers who are visually impaired may have trouble perceiving a diagram, even if they have the knowledge that is supposed to be tested by the item based on the diagram. Test takers with hand injuries may be unable to use an answer sheet or manipulate a computer input device.

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

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

this case, so the drawings are not essential. The drawings, however, are helpful as stimuli when it cannot be assumed that the test takers share a common native language. A judgment must be made about whether or not the advantages of the helpful aspects for some test takers outweigh the disadvantages for other test takers. Furthermore, accommodations should be made for test takers with disabilities who are disadvantaged by the items.

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

Examples of physical barriers. The following are examples of physical barriers in items or stimuli that may be unnecessarily difficult for test takers, particularly for people with certain disabilities. ETS test developers are told to avoid these barriers, or others like them, if they are neither essential nor helpful for measuring the intended knowledge or skill.

- unnecessary use of visual stimuli (e.g., charts, diagrams, graphs, and maps that are not important for validity); visual stimuli that are more complex, cluttered, or crowded than necessary;
- fine distinctions of shading or color to mark important differences;
- lines of text that are vertical, slanted, curved, or anything other than horizontal; fonts that are hard to read; or text that does not contrast sharply with the background;
- letters that look alike or sound alike used as labels for different things in the same item or stimulus;
- special symbols (unless that is standard notation in the tested subject, such as Σ in statistical notation).

In addition, ensure that audio presentations are clear enough to avoid having the quality of the audio serve as a source of difficulty. Similarly, text and images displayed on a computer screen should be clear enough to avoid having the quality of the display serve as a source of difficulty. Reduce the need to scroll to access parts of stimulus material or items to the extent possible, unless the ability to scroll is being measured.

Physical barriers are likely to be very similar across countries because they are caused by sensory and motor problems that can affect any human being rather than by cultural, linguistic, or other issues that vary across countries. What physical barriers are of concern in your country?

ESTABLISH PROCEDURES

The fairness guidelines should be accompanied by detailed instructions for their use. ETS has established procedures that it applies to all fairness reviews. Those procedures are listed below.

Item writers and reviewers should be trained to follow the fairness guidelines.

All items and stimuli should be reviewed for fairness by trained fairness reviewers before being 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 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., videos, extended reading passages, technology enhanced items, simulations) should receive a fairness review before any substantive work is done. The early 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.

For use in your country, which procedures should be adopted? Which modified? Which rejected? Are additional procedures required?

TRAIN USERS OF THE GUIDELINES

People who will be involved in the test-development process in your country must be trained to use your fairness guidelines. In the experience of ETS, simply reading the guidelines document is not sufficient. The people who will write and review 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. Items that are unfair in obvious ways are helpful only in the earliest stages of training. Items that cause disagreements and discussions among the trainees are the most useful training items.

Training should stress that for many of the guidelines, compliance is a matter of degree rather than a clear binary decision. At what point does the difficulty of language become a an irrelevant barrier to performance? How controversial does material have to be to violate a guideline? Material that seems acceptable to some reviewers may be rejected by other reviewers. How important for validity does content have to be to justify its inclusion if it appears to be out of compliance with a guideline? Subject-matter experts may disagree about the importance of certain content. Judgment is required to interpret the guidelines appropriately.

An overly lax interpretation of the guidelines may allow unfair content into your tests. An overly rigid application of the guidelines may be harmful as well, by interfering with validity and authenticity. Therefore, the individual guidelines must be applied conscientiously, but with an awareness of the need to measure important aspects of the intended content with realistic material that is appropriate for the test-taking population.

ETS test developers are told to consider the following factors in deciding whether or not material is in compliance with the guidelines.

Types of tests. In the application of fairness 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.

Age, sophistication, and previous experience of test takers. In general, the older and more sophisticated the test takers, the more liberally the guidelines should be interpreted. Also, consider the kinds of material that test takers are likely to have been exposed to in deciding whether some test material is likely to offend or upset them. Brief prior exposure does not necessarily justify the inclusion of upsetting material in a test. Furthermore, not everything that is discussed in class with the guidance and support of a teacher and the opportunity to ask questions is necessarily appropriate in a test. Unlike content discussed in class, content in a test may be an irrelevant barrier to performance and may possibly have negative consequences for the test taker. If, however, test takers have become accustomed to the material through repeated exposure in their studies, their occupations, or their daily lives, it is not likely that encountering it again in a test would be excessively problematic.

Directness of the material. Items and stimuli about innocuous topics are generally acceptable, even if a scenario could be constructed in which they might possibly be upsetting for some test takers who had undergone a particular experience. Contexts that directly mention an upsetting experience are less likely to be acceptable. For example, a mathematics item about the average speed of a car should not be

construed as potentially upsetting for test takers who have been involved in a car accident. On the other hand, a mathematics item about the average number of children killed per year in car accidents would be unacceptable, unless it were important for validity and no equally important substitute were available.

Extent of the material. A brief mention of a problematic topic may be acceptable even though a more extended, detailed discussion of the topic should be avoided. For example, a statement that a cat killed a wild bird might be acceptable, but a graphic description of how the cat caught, toyed with, and eventually clawed apart a baby robin would probably not be acceptable.

Which of these factors should be included in the training of test developers in your country? Should any factors be added?

MONITOR AND REVISE GUIDELINES

Keep in mind that it is impossible to develop rules and examples for fairness in assessment 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 useful to have some person or group be responsible for the upkeep of the fairness guidelines. Document disagreements between item writers and fairness reviewers. If the same disputes keep occurring, that is an indication of some ambiguity or lack of coverage in the written guidelines that should be remedied.

It is a good practice to schedule a review and revision of your fairness review guidelines every five years or so. Events that change perceptions of fairness may require more frequent revisions of the guidelines.

ADDITIONAL FAIRNESS ACTIONS

As important as they are, fairness guidelines by themselves are insufficient to ensure that the testing process is fair. We recommend using the following methods of enhancing fairness in addition to the fairness guidelines.

Treat all test takers respectfully and impartially. Fairness requires that all test takers be treated with respect and without regard to irrelevant personal characteristics such as race or ethnicity. This holds

true throughout the testing experience. Strive to give all test takers respectful treatment, equal access to relevant testing services, and useful information about the test. As part of treating test takers appropriately, create test preparation materials and tests in formats accessible to persons with disabilities. Provide needed accommodations for test takers with disabilities so that the test measures relevant knowledge and skills rather than the irrelevant effects of a person's disability.

Use diverse contributors. It is useful to obtain contributions to tests from people who represent relevant perspectives and diverse groups. Representatives of various groups can be included among the people who determine the knowledge, skills, and other attributes to be tested. Additional means of obtaining contributions to help maintain fairness include involving people who are members of various racial and ethnic groups as item writers and reviewers, as test reviewers, and as essay scorers.

Use Differential Item Functioning (DIF) Statistics. Consider using statistical measures of differential item functioning (DIF) as an empirical check on the fairness of items. DIF occurs when people in different groups perform in substantially different ways on a test item, even though the people have been matched in terms of their relevant knowledge and skill as measured by the test. The statistics are applied whenever the data would be useful and sample sizes are large enough to allow meaningful results. If DIF data are available, assemble tests using items with low DIF to the extent possible. If data are unavailable at assembly, DIF can be calculated after test administration, but before operational scoring. Items with high DIF can be reviewed for fairness by people who have no vested interest in the test. Any items judged to be unfair can be removed before the test is scored. See Dorans (1989) and Zieky (1993) for more information about calculating DIF and about using DIF appropriately.

Obtain validation information. A crucial aspect of fairness is validation. Essentially, validation is the collection of evidence to evaluate the extent to which the inferences made on the basis of test scores are appropriate. Multiple lines of evidence are pursued in validation efforts. Some important aspects of validation are, for example, demonstrating that the people who determined the specifications for the test had the training and experience necessary to do a competent job; showing that the different parts of the test relate to one another and to external criteria as theory would predict; and determining the extent to

which the items sample only relevant knowledge and skills. See Messick (1989) and Kane (2006, 2013) for more information about validity.

Specify appropriate test interpretation and use. Even a fair test can be used unfairly. For example, when the opportunity to learn the tested material is not equally distributed, interpreting scores as measures of innate ability is unfair. Specify the appropriate interpretation and use of your tests and make the information available to score recipients. Investigate plausible allegations of misuse and inform the score recipient how to use the test appropriately.

Survey relevant research on fairness. This manual is intended to focus on the fairness of test content in different countries. For other aspects of fairness in testing such as preparation, administration, analysis, scoring, and reporting, you must consult other sources. The publications listed in the References section of this manual will be helpful**.** In addition**,** ETS supports a great deal of research directly related to test fairness. Free reports of the research are available to you at www.ets.org. Click on the tab for Research and use the search facility to find relevant documents.

CONCLUSION

Though there are universal principles for fairness in assessment, the principles will be more effectively applied if they are augmented by specific guidelines tailored for the country in which the test will be used. The development of clear and specific fairness 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. Problems in using the guidelines should be documented and the guidelines revised to alleviate the problems and to remain current. The results of your efforts will be fairer and more valid tests for all of the people who take them and for all of the people who make use of the scores.

Some Useful References on Fairness in Assessment

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

American Psychological Association. 2010. *Publication manual of the American Psychological Association.* Washington, DC: Author.

Bartram, D. (2001). International guidelines for test use. *International Journal of Testing, 1*(2), 93-114.

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

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

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

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

Educational Testing Service. (2015). *ETS guidelines for fair tests and communications.* Princeton, NJ: Author.

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

International Language Testing Association. (2007). *Guidelines for practice*. Author. Retrieved from http://www.iltaonline.com/index.php/enUS/component/content/article?id=122

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

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

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

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

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

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

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

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

Measuring the Power of Learning.®

www.ets.org