

MASTERING ASSESSMENT: A SELF-SERVICE SYSTEM FOR EDUCATORS

Assessment Bias: How to Banish It

Second Edition



W. James Popham

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editor in Chief: Paul A. Smith
Senior Marketing Manager: Danae April
Marketing Assistant: Elizabeth Mackenzie-Lamb
Senior Managing Editor: Elaine Ober
Manufacturing Buyer: Megan Cochran
Full-Service Project Management: Lynda Griffiths
Composition: TexTech, Inc.

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on appropriate page within text (or on page 32).

Copyright © 2012 Pearson Education, Inc., publishing as Allyn & Bacon, 501 Boylston Street, Boston, MA, 02116. All rights reserved. Manufactured in the United States of America. This publication is protected by Copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. To obtain permission(s) to use material from this work, please submit a written request to Pearson Education, Inc., Permissions Department, 501 Boylston Street, Boston, MA 02116, or email permissionsus@pearson.com.

10 9 8 7 6 5 4 3 2 1



www.allynbaconmerrill.com

ISBN 10: 0-13-273-4907
ISBN 13: 978-0-13-273-4905





PREFACE

This booklet, one in a series of 15 booklets contained in *Mastering Assessment: A Self-Service System for Educators (MA)*, was written chiefly for those who currently teach in our schools. Because I was once a public school teacher, I have lasting respect for our nation's teachers. Thus, I have not tried to write a series of tiny *textbooks* wherein an author attempts to teach a reader. In this booklet, therefore, you will find no practice exercises, no end-of-chapter quizzes, and no interminable list of additional references. Instead, I tried to write each booklet in the form of a colleague-to-colleague conversation—a somewhat one-sided conversation to be sure. What I attempted to do in every booklet was explain some assessment ideas I regard as important for today's teachers to understand. I attempted to provide those explanations employing the language I'd use if I were sitting with a fellow teacher in a faculty lounge. Even so, you'll occasionally encounter a dollop or two of irreverent whimsy in these *MA* booklets because I seem altogether incapable of controlling my silliness propensities—in faculty lounges or elsewhere. Given my hope that readers will acquire several assessment-related *understandings* from each *MA* booklet, in the spirit of collegial candor I have laid out those anticipated understandings at the start of every booklet. Moreover, at each booklet's conclusion I have briefly reiterated what I regard as the essence of those understandings. After all, even colleagues sometimes need to repeat themselves.

WJP

Anticipated Understandings

After reading this booklet, you should understand:

-  *What constitutes assessment bias and the two ways such bias can reduce the validity of test-based inferences.*
-  *That “disparate impact” of an educational test does not automatically signify the presence of assessment bias.*
-  *The nature of three common sources of assessment bias: racial/ethnic bias, gender bias, and socio-economic bias.*
-  *How assessment bias can be reduced in both large-scale tests and classroom tests.*

Bias is a bummer. It’s definitely a bad thing. However, if you prefer a less colloquial description of the badness and bum-merness of bias, you’ll probably find a dictionary definition somewhat more satisfying:

bias: prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair. *The New Oxford American Dictionary*, New York: Oxford University Press, 2001.

Bias, as you can see, is not only prejudicial, it’s often unfair. Biased people, because of their prejudgments, often make bad decisions about persons, groups, or things.

Well, if bias is a bad thing in general, be assured that bias is every bit as vile when it worms its way into educational measurement devices. *Assessment bias*, therefore, also reeks of full-blown badness.

THE NATURE OF ASSESSMENT BIAS

Educational tests, if they’re good ones and if they’re used properly, permit us to make accurate inferences about the skills and knowledge students possess. Because those skills and knowledge can’t actually be seen, because they are *covert*, we rely on students’ *overt* performances on educational tests to arrive at interpretations about what it is that students know and can do. The process of making test-based inferences about students, in fact, represents the bedrock of educational assessment. If teachers’ inferences about their students are accurate, then teachers can make appropriate decisions about the best ways to instruct those students. On the other hand, if teachers’ inferences about students are inaccurate, then the instructional decisions those teachers make are likely to be unsound.

The accuracy or, more technically, the *validity* of test-based inferences regarding students' skills and knowledge is therefore all-important if students are going to receive first-rate instruction. And that's where assessment bias, in all its badness, comes bouncing into our backyard. Assessment bias meaningfully mucks up the accuracy of measurement-based interpretations about students. Assessment bias, in other words, diminishes the validity of educators' test-based inferences about students. And it is for this reason, of course, that teachers need to understand the nature of assessment bias. The more you know about a villain's characteristics, the better will be your chances of vanquishing that villain. Here, then, for purposes of villain detection, is a formal definition of assessment bias:

Assessment bias is present whenever one or more items on a test offend or unfairly penalize students because of those students' personal characteristics such as race, gender, socioeconomic status, or religion.

As you can see from this definition, there are two aspects of a test's items that can contribute to assessment bias: *offensiveness* and *unfair penalization*. Let's look at both of those nasty notions right now.

Offensiveness

A test item is apt to offend a particular group of students when the item's content somehow denigrates the particular group to which those students belong. For example, if an item on an important exam happened to include a disparaging remark about Mexicans, and both parents of a student who was taking the exam had been born in Mexico, it is obvious that the student would most likely be offended. Moreover, common sense tells us that students who have been offended aren't likely to perform optimally—not only in responding to the offensive item but also in responding to subsequently encountered items on that test.

Why on earth, you might ask, would a teacher include a disparaging remark about any group of human beings in a test item? Well, no teacher I've ever met actually sets out to create test items that deliberately denigrate individuals or groups. Yet, from the student's perspective, indeliberate denigration is every bit as offensive as deliberate denigration. Offensive content in test items can distract, can rile, and can hurt. However, I believe that most teachers toss offensive content into their test items without recognizing it. That is, they do so unwittingly.

Assessment Bias: The presence of one or more items in a test that offend or unfairly penalize students because of students' personal characteristics

I subscribe to Howard Gardner's view of multiple intelligences, in which he argues that people possess differing degrees of verbal smarts, quantitative smarts, interpersonal smarts, and so on. Well, if it's true that certain individuals possess particularly sensitive antennae regarding other people's feelings, then the interpersonal antennae of some other individuals unfortunately seem to be coated with cotton. Putting it plainly, some folks are really rotten at figuring out how other people are apt to react to, for example, oral or written comments. Some teachers are simply unaware that the content of a test item might be offensive to particular groups of students—for instance, females, Native Americans, or Muslims. But, as I just indicated, a test item can offend students even if the item was written by an interpersonally insensitive, albeit well-meaning teacher.

Offensiveness: Test items containing content that insults, irritates, or causes pain to students because of those students' personal characteristics

Later on in the booklet, we'll look at procedures to systematically spot offensiveness in test items, but let me jump

that content's starting gun a bit by indicating that the very best way to identify potential offensiveness in your own test items is to become extraordinarily attentive to the offensiveness potential of any test item you write. That's it. Simply acquire gobs of *extraordinary attentiveness* to whether what's in an item might rub some students the wrong way. This is, of course, far easier to recommend than to implement. You don't turn an interpersonally insensitive clod into an interpersonally perceptive one merely by asking—even by asking politely.

If a teacher's class contains a large number of students from a particular subgroup—for example, children from Somalia or children who are physically disabled—it's fairly easy for that teacher to be attentive to what might offend those particular youngsters. They're so obviously present in the class. But the trick is to focus on *every* student, not only on large, in-your-face subgroups. If there's one Jewish student in your class whom a test item offends because of its anti-Semitic content, that's one student too many. If there's one Vietnamese student in your class whom a test item offends because of its Asian-denigration content, that's one student too many.

Students don't have to be Jewish or Vietnamese themselves to be offended when items on a test contain anti-Semitic or anti-Asian content. Clearly, many students will be—and should be—offended by test items that disparage any group.

Unfair Penalization

A second sort of assessment bias can arise when a test item actually penalizes a student because of that student's personal characteristics such as gender or geographic locale. Let me use an example to show you how an item's content can disadvantage students who are members of a particular group, in this example, female students.

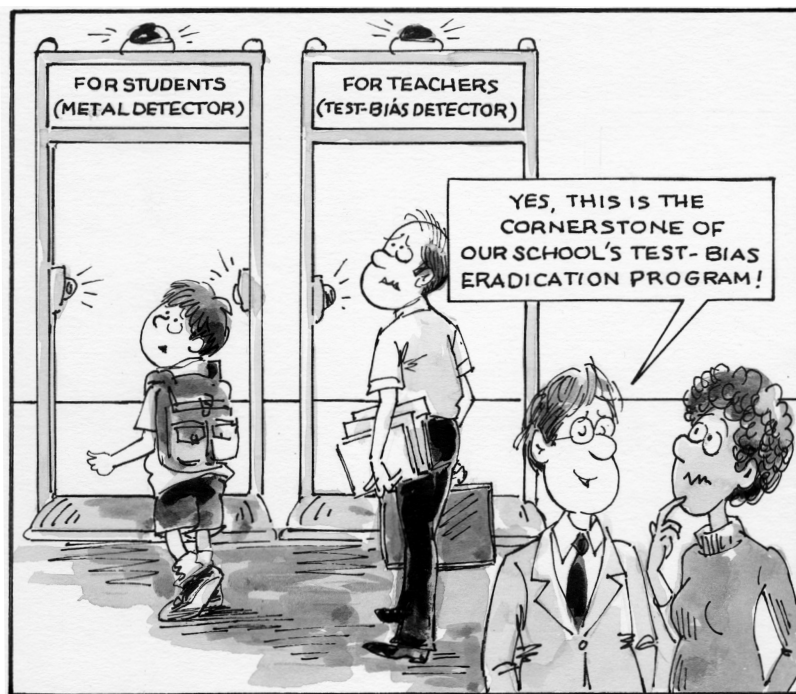
Let's say you're a middle school math teacher and you've put together an end-of-unit exam intended to see how well your students can solve mathematical word problems. One of your test's first few items is set in the context of a professional football game, and students are supposed to solve a word problem that includes a key reference to YAC. In order to solve the problem, students definitely need to understand that YAC refers to "yards after catch." That is, when a football is thrown to a receiver, and the receiver catches the ball, how many additional yards does the receiver gain before being tackled by someone from the other team?

Although female students as a group, with every passing year, learn more and more about all sorts of sports, and although televised sports events now clearly and consciously include female analysts and announcers, the fact remains that more boys than girls are likely to understand the meaning of YAC. Not every boy is a follower of the National Football League,

of course, and not every boy will know the meaning of YAC. Similarly, some girls will surely know what YAC means. Perhaps if there were tackle-football teams for girls in high school and college, this wouldn't be the case, but that is the subject of another booklet altogether! In the meantime, because far more boys than girls will know what YAC signifies, then more boys than girls will almost certainly be able to answer the professional football math problem correctly. Girls will be unfairly penalized because it was the item's gender-linked content that disadvantaged female students, not the mathematics involved.

Just as we saw that an offended student will underperform not only on an offending item but on other items in the test, unfair penalization works in essentially the same way. If a student encounters an item containing key content that the student regards as incomprehensible, then the student may feel inadequate not only with respect to the particular item involved but also with respect to other items on the test. Unfair penalization, as its name suggests, is unfair.

Just as with the detection of offensiveness in test items, the very best way for teachers to spot items whose content might unfairly penalize some students is to maintain an attitude of *extraordinary attentiveness* to the possibility of unfair penalization. In a sense, you are trying to make sure that every one of your students will be in a position to respond



correctly to a test item—assuming that the student has put in the necessary effort.

And this brings up a point that needs to be mentioned about the adjective *unfair* in the phrase *unfair penalization*. Not all of your students will earn perfect scores on your tests. Some students, indeed, will perform miserably on certain of your tests. Let's assume, however, that you've done a terrific job of teaching something, but that one of your students (I'll call him Lazy

Larry) really paid little attention in class and, to worsen matters, never completed several of your key homework assignments. When the class is tested, Lazy Larry predictably flops—big time. His low test score earns him a solid F on the test. And that F, of course, is Lazy Larry's penalty. Now, was this penalization unfair? Of course it wasn't. Lazy Larry, the lout, richly deserved his test-based F. For assessment bias to be present, the student's penalty must be *unfair*. That is, it must be a student's personal characteristics that made it difficult, if not impossible, for the student to perform well on one or more of a test's items.

Unfair Penalization: Test items containing content that unjustly prevents one or more subgroups of students from performing well because of those students' personal characteristics

Inference Distortion

Okay, you've now seen that the two contributors to assessment bias are test items' offensiveness and unfair penalization. What's the impact of assessment bias if it actually is present in an educational test? Well, and this is the nub of the nastiness flowing from assessment bias, tests that are biased will produce students' scores from which valid inference will rarely be made.

Just think about it a bit. If a particular subgroup of students has flopped on a section of a biology test because they've been taught in their religion (both learned in church and learned from family members) to reject the accuracy of certain biological content, the reason for their poor performance on the biology test may not be because they don't know what's being tested. Rather, they may know very well what's being tested—but may simply not accept it. A teacher's inference about those students' knowledge about biology is, therefore, likely to be invalid.

Similarly, remember the previous example about a mathematics item nestled in a football context where students needed to know what YAC meant. Well, if that biased item caused many of a teacher's female students to perform less well on a test than they would have if the teacher had yanked YAC from the item, any test-based inference about the female students' math understandings will probably underestimate those understandings.

Because accurate test-based inferences usually contribute to better instructional decisions by teachers, the more assessment bias there is in an educational test, the less the likelihood of teachers arriving at accurate test-based inferences about students' skills and knowledge. Consequently, kids will be less well taught if teachers base their instructional decisions on the results of biased tests. Bottom line: Eradicating assessment bias is in the best interest of both the students *and* the teacher.

DISPARATE IMPACT: A CLUE, NOT A VERDICT


A measurement misconception held by many educators is the following: If a test has a discernible *disparate impact* on certain subgroups of students, then the test is biased. That's simply not so. But teachers need to understand why disparate impact and assessment bias are not equivalent concepts.

Let me introduce this topic by calling on a personal experience of some years ago. At the time, I was directing a test-development group whose mission it was to create statewide achievement tests. (I hasten to add that, for more than a decade now, I haven't been involved in such projects. Indeed, I am a "lapsed test developer.") Anyway, my test-development group was building a set of high-stakes tests for a southern state. During the development process, we had created a whole flock of test items for students at various grade levels, and had then tried out those items as part of a large-scale field test. The incident I want to tell you about occurred when a committee of about 30 teachers reviewed the field-test results in order to decide which items should subsequently be used in the operational statewide tests.

The largest group of minority students in this state consisted of African American children, and we consequently

had made sure that at least half of the members of our item-review committee were African American educators. As the committee's members went through the items, one by one, we presented to them the results of students' field-test performances—including each item's difficulty level. We also presented differences, if any, between the difficulty levels for white and African American children on all items. If there was any item for which a 10 percent differential existed between the performances (difficulty levels) of those two racial groups, we flagged the item so the committee could decide whether the item was biased and should therefore be discarded from the item pool.

I had not previously moderated those sorts of item review meetings and, with a desire to bend over backward not to include biased items in the state's new tests, I was quite willing to chuck all of the flagged items on which African American children had not done as well as their white counterparts. And, indeed, the committee spotted some items that, because those items might offend or unfairly penalize African American students, were tossed out.

 **Disparate Impact: When an item on a test results in substantial differences in the success rates of separate subgroups of students**

But, and here's where I personally learned an important lesson, the vast majority of flagged items showing disparities, sometimes substantial ones, in favor of white students, were *retained*—not discarded. The items were retained because the African American teachers on the committee, time and again, made comments such as the following: “This flagged item covers content that *all* of our state's students definitely need to learn. The poor performance of African American children during the field test lets us know that those children hadn't been properly taught. If you delete this item, we'll be unable to determine if that situation has changed. Don't you *dare* remove this item from the item pool!”

Clearly, the item-review committee recognized that merely because there's a difference in the performance of certain groups of students on a test item, this does not automatically render the item biased. Disparate impact of an item, or of an entire test, may indeed reflect bias. But it may also reflect prior instructional inadequacies. Those African American teachers taught me an important lesson: If an item measures something children need to know, and there's nothing in the item that might offend or unfairly penalize students, then the item needs to remain in a test. Flagging a disparate-impact item for further review is okay; automatic and forcible removal of the item isn't. Happily, in recent years most of the large test-development firms in the United States seem to understand the important

distinction between a test item that yields a disparity in subgroup performances and a test item containing content that offends or unfairly penalizes certain subgroups.

Later we'll review the technical tools that are now used to detect disparate impact in a test's items, but let's start by disabusing ourselves, once and for all, of the idea that disparate impact and assessment bias are coterminous. (Coterminous, meaning "having equal boundaries," is one of my favorite words—I try to work it into my assessment-related writing whenever the occasion permits.) Disparate impact is a great way to get an item considered more carefully for the possibility of bias. Disparate impact, as I hope you see, definitely does not guarantee the presence of bias.

THREE COMMON SOURCES OF ASSESSMENT BIAS

Now, in an attempt to lay out the sorts of bias that teachers must be on guard against, I want to deal with the three sources of bias most frequently found in educational tests. I'm going to describe briefly the nature of each of these bias sources, then give you two examples of test items in which there's bias attributable to the particular source I've been describing.

Before doing so, however, I need to draw on the colleague-to-colleague nature of this and the other *MA* booklets.

More specifically, I want to make a brief confession. This confession, simply put, is that I've always found it difficult, without embarrassment, to devise examples of test items that are blatantly offensive. For example, suppose I cranked out a fictitious test item that harshly disparages Hispanic Americans in a manner apt to be hurtful to Hispanic American children (or adults). Even though the fictitious item is only supposed to be an illustration, and an illustration that's explicitly intended to help teachers avoid the kind of offensive content contained in the item, I am still uneasy about having such items even appear in print—especially in something I've written. So, if you encounter content in any of the next six items that you regard as offensive, please remember that this itself is the point: I'm including these items for illustrative purposes. Even though that's so, I'm still embarrassed when I do so.

Racial/Ethnic Bias

Description. In most minds of most people, when anyone talks about test bias, it's usually thought that such bias is linked to students' race or ethnicity. Moreover, because *racial/ethnic bias* is widely acknowledged to exist in educational tests, and is so fundamentally wrong, I want to lead off with a consideration of what's involved in that source of assessment bias.

First, let's do a bit of term-tidying. Without going into an elaborate sociological semantic analysis, here's what I mean when I refer to a student's *race* or to a student's *ethnicity*. I accept my *Oxford American Dictionary's* definition of *race* as "each of the major divisions of humankind, having distinct physical characteristics" (*The New Oxford American Dictionary*, New York: Oxford University Press, 2001). Turning to *ethnicity*, again I defer to my dictionary (any dictionary with 2,023 pages in it should warrant at least some deference), which defines *ethnicity* as "the fact or state of belonging to a social group that has a common national or cultural tradition" (*The New Oxford American Dictionary*, New York: Oxford University Press, 2001). So, in general terms, *racial* refers to a major human species subdivision, but *ethnic* refers to a subgroup within a race. So, for instance, if we chose to designate "Asian" as a race, then we could surely identify different Asian ethnic groups because of national traditions (Vietnamese versus Korean) or cultural traditions (Taiwanese Chinese versus San Francisco Chinese). But this straightforward distinction often becomes blurred, because even my beloved and hefty *Oxford American Dictionary* also turns out to define *race* as "an ethnic group." So much for my efforts to do some demarcation drawing!

In any event, then, that's the way I'm going to be referring to race and ethnicity. *Racial* will refer to a major anthro-

A Dictionary Discreditation

"Although ideas of race are centuries old, it was not until the nineteenth century that attempts to systematize racial divisions were made. . . . Theories of race asserting a link between racial type and intelligence are now discredited. Scientifically it is accepted as obvious that there are subdivisions of the human species, but it is also clear that genetic variation between individuals of the same race can be as great as that between members of different races."

The New Oxford American Dictionary

New York: Oxford University Press, 2001, p. 1402

pological division of humankind—for instance, Asian. *Ethnic* will refer to subgroups within those divisions—for example, San Francisco Asians of Chinese descent versus Los Angeles Asians of Cambodian descent. To regard different ethnic groups drawn from the same racial group as being equivalent to one another is surely naïve.

However, when a teacher tries to avoid any content in test items that will offend or unfairly penalize students because of their race or ethnicity, this teacher is usually being

attentive to the same sorts of concerns. Accordingly, that's why I've decided to use the descriptor "racial/ethnic" to describe this first source of assessment bias. If you want to fuss with whether an offensive segment of a test item is more focused on race than on ethnicity, go to it. To my mind, if an item offends or unfairly penalizes a student because of the student's race or ethnicity, it really doesn't matter all that much which of those two factors gets the blame. The biased item either has to be fixed or, perhaps, sent to a suitable paper shredder.

Illustrative items. Let me show you, now, a pair of items that suffer from racial/ethnic bias. Please consider the test item in Figure 1 which, as you'll see, is a mathematical item attempting to see whether students understand the meaning of three mathematical symbols: $>$ (greater than), $<$ (less than), or $=$ (equal to). Note that the item is set in the context of General George Custer's defeat at the battle of the Little Bighorn.

The illustrative item in Figure 1 is loaded with racial/ethnic bias. Let's start with what the item is supposedly attempting to do—namely, measure whether students understand how to use three mathematical symbols. Well, if a student was born and raised in the United States, that student would most likely have learned at some time that General Custer's forces had been wiped out by members of the Lakota and Cheyenne tribes. And, if the student happened to know that fact, then

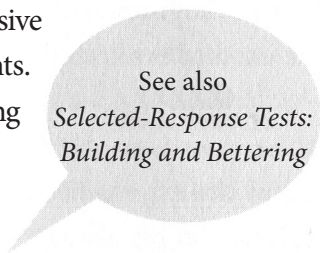
Figure 1. An Illustrative Mathematics Item Embodying Racial/Ethnic Assessment Bias

In 1876, General George Armstrong Custer and his troops fought Lakota and Cheyenne warriors at the Battle of the Little Bighorn. If there had been a scoreboard on hand, at the end of that battle which of the following scoreboard representations would have been most accurate?

- A. White Men $>$ Red Men
- B. White Men $=$ Red Men
- C. White Men $<$ Red Men
- D. All of the above scoreboards are equally accurate

Choice C (White Men $<$ Red Men) would have probably been the student's choice. Choice C appears to be the item's correct answer. However, if a student had lived in the United States for only a few years, having emigrated from another nation, then the student might never have heard about Custer's defeat. And, without knowledge of that historical event, it would be impossible for the student to answer the item correctly—except by guessing. The item unfairly penalizes any racial/ethnic group of students who aren't familiar with General Custer's demise.

Beyond that fundamental shortcoming, of course, there is the offensive contrast between “White Men” and “Red Men.” Surely such language would prove abrasive today to many Native American students. Furthermore, the whole premise of having a scoreboard to represent the proportions of human beings who were killed in battle is staggeringly insensitive.



See also
*Selected-Response Tests:
Building and Bettering*

Finally, Choice D (“All of the above scoreboards are equally accurate”) should never be used as an option in multiple-choice items. The illustrative item in Figure 1 is, in short, ghastly. (Can you see why I get queasy about using such awful items, even for illustrative purposes?)

Let’s consider another item that reeks of racial/ethnic bias. Please look at the item in Figure 2 dealing with the history of Hawaii. As you can see, the item asks students to select the group of immigrant workers who moved most slowly in making economic strides. I suppose that it might be possible to come up with a correct answer for this repugnant item. (I certainly don’t know what a correct answer would be.) However, imagine that you are a child in one of Hawaii’s public schools and you encountered this item on a test. Suppose, further, that you’re a Portuguese student, a Filipino student, a Chinese student, or a Japanese student, and this item makes

Figure 2. An Illustrative History Item Embodying Racial/Ethnic Assessment Bias

Hawaii has been influenced greatly by immigrant workers from other nations. These groups originally toiled as low-cost laborers on Hawaii’s numerous pineapple and sugar plantations. Subsequent generations of these immigrant groups, however, entered all phases of Hawaiian society. Which of the following groups was the *slowest* to make economic progress in Hawaii?

- A. Chinese
- B. Japanese
- C. Portuguese
- D. Filipino

you think about your ethnic background—and may make you feel bad about it. Would those negative feelings possibly mess up your performance when responding to subsequent items in the test? I’d think so.

This illustrative item is apt to offend some students—from any of the four ethnic groups seen in the item’s answer choices—because the implication of the item is that one of the

immigrant groups is “slowest.” Children don’t like to think of themselves as being in a slow group. Members of a “slow” group are not thought to be as able as are members of other groups. This absurd item is laden with potential offensiveness to children.

English-language learners. Just briefly, I’d like to consider assessment bias insofar as it pertains to English-language learners. I’m thinking of children who have arrived in the United States having grown up in a country where a language other than English is the dominant language. Clearly, if we were to administer an English-language test to such students within weeks of their arrival in an English-speaking nation, irrespective of whether the test dealt with science, mathematics, or language arts, those recently arrived students would most likely flounder on the test. It’s blinking tough to respond appropriately to a test item that, to your eyes, is made up of indecipherable gibberish. Let’s concede that such recently arrived students’ likely failure on an English-language test would constitute a penalty. But is that penalty unfair?

Remember, assessment bias scampers onto the scene only when items offend or *unfairly* penalize students. But it surely seems unfair, at least to me, if third-graders from Honduras are forced to endure a series of English-language exams only weeks after arriving in Nebraska. Not only will

such an assessment have a negative impact on those third-graders’ self-esteem but, in most instances, the test’s results will lead to invalid inferences about the students’ actual skills and knowledge. However, here’s why I cunningly qualified the previous sentence by tossing in the squishy prepositional phrase, “in most instances.”

If one of the tests is a mathematics exam, and we’re trying to get an accurate fix on these Spanish-speaking children’s mastery of certain mathematics content, then their low performance on an English-language math test is probably *not* indicative of their actual mathematics capabilities. The students can’t display their math prowess accurately because they can’t understand the English-language lead-ins to the test’s items. The same situation would also prevail in the case of a science exam. Inaccurate inferences about children’s mastery of science content will be made because of the language used in the science test.

However, and here’s where you need to think carefully about an important issue, if an English-language test is administered to recently arrived Spanish-speaking children measuring those children’s ability to read and write *in English*, will the children’s poor performances on that test constitute an *unfair* penalty?

To make this issue more graphic, let’s think about just the *reading* portion of a language arts test. If teachers want to

find out if a Spanish-speaking child can read English-language materials, and the student cannot do so when taking an English-language test, won't the teacher's inference that "the child can't read in English" be accurate? Sure it will. Unlike a math test where the focus is on the student's math competence, and English is merely the exam's delivery vehicle, an English-language math test is likely to result in invalid inferences about students' mathematical mastery. Yet, in the case of language arts, say, reading, where English is not only the exam's delivery vehicle but also the substance of the exam itself, inadequate performances by English-language learners provide teachers with the evidence they'll need in order to conclude—accurately—that this child can't read in English.

Where *unfairness* raises its nasty noggin in this scenario, at least in my opinion, revolves around whether English-language learners have been given a reasonable amount of English-language instruction so that they have had a decent chance to learn what's being measured. Incidentally, here is another instance where colleagues can disagree, because my stance on this issue is surely not a universally held one. But I regard it as wrong, inhumane, and unjustifiable to test English-language learners before they have had sufficient instruction in English to have a genuine opportunity to succeed. I'm not at all sure about the necessary duration of

English-language instructional time that certain groups of students would need. But I'm pretty certain, based on my conversations with many teachers of English-language learners, that only one year of instruction in English is frequently insufficient. Yet, I'll leave that call to the teachers involved, for I'm sure that there must be meaningful differences in the necessary language transition times when diverse language such as Farsi, Portuguese, Amharic, and Tagalog are involved.

If English-language learners are given a reasonable amount of instruction in English, and I'm thinking along the lines of two or three years, I believe the use of English-language tests will be apt to yield valid inferences because those tests will, at that point, be free of assessment bias. There will, by that time, be no unfair penalization of the English-language learners involved.

Let's turn, now, to another common source of assessment bias, a source that periodically gets its share of attention from the media. I'm referring to the kind of bias that hinges on whether the students being tested are boy students or girl students.

Gender Bias

Description. Gender bias takes place in educational tests when a student is offended or unfairly penalized because the student

is a male or a female. It seems, given the historical dominance of males in most settings, that gender bias in educational tests is most likely to have a negative impact on females. Women have, over the years, been relegated to lower status, lower salaries, and often lower esteem in most of this world's societies. Naturally, then, that ubiquitous societal bias against females sometimes sneaks its way into the items on educational tests.

For instance, in the early decades of the twentieth century, at least in Western civilizations, a woman's place was widely thought to be "in the home," bearing children and caring for them. It was the role of men to be the family's "breadwinner," and to seek success in the world outside the family. Even today, some of the residue of such patently unequal expectations for males and females remains with us. And although women have made striking gains on many fronts during the last several decades, some of those earlier notions about the "proper" role of females can sometimes be found in items on very important educational tests.

But gender bias chops in both directions. It is certainly possible to create test items that are biased against males. To illustrate, in current society it seems clear that females are, in general, more attuned to style and fashion than are males. I know there are exceptions, of course, and that some men are extremely fashion conscious—perhaps even more so than most

women. But when we deal with assessment bias, and its sources, we need to be thinking about *groups*, not about the atypical members of such groups. Thus, suppose a test question on a high school exit exam was set in a context where familiarity with current fashion and style in women's attire was needed for students to answer the item correctly. If 100 high school boys and 100 high school girls tackled that test item, we'd be likely to see far more of the girls succeed on the item than boys. The probabilities of success would clearly favor girls. In other words, girls would have an edge over boys on the item because girls would be more likely to be conversant with the item's content. And boys, therefore, would suffer the consequences of assessment bias because the item would unfairly penalize them. Gender bias, just like racial/ethnic bias, is a bad thing.

Illustrative items. Let's look, now, at a pair of items that display gender bias. First, please look at the item in Figure 3 which, as you can see, is a mathematics item dealing with executive-level salaries for men and women in four large companies. To answer the item correctly, a student will need to decide how best to compare the four salary ranges that have been given for male and female executives. One reasonable way of doing so might be simply to identify the midpoint of each salary range, then compare the differences between the

Figure 3. An Illustrative Mathematics Item Embodying Gender Assessment Bias

Recently, in end-of-year reports, four of the world’s largest companies released annual salaries of male and female executives. Please review the boxed information below, then select the company in which the difference in male–female executive salaries appears to be the smallest. Circle the letter of your answer below the box.

	Annual Salaries Ranges—in Thousands of Dollars			
	Company A	Company B	Company C	Company D
Male Executives	70–110	80–120	90–150	90–190
Female Executives	60–100	60–80	70–90	100–120

A. Company A

B. Company B

C. Company C

D. Company D

annual salaries of males and females as represented by those midpoints. Using that approach, it turns out that Company A has the smallest gap between male and female executives—that is, between the midpoints of the salary ranges given in the test item. Company A, incidentally, is also the company where the lowest salaries are paid to executives.

Corsage Bias

When I was a high school student, junior prom dances were big deals. And I still vividly recall an instance as a teenager when gender bias caused me some serious

embarrassment. At that stage of my life, everything seemed serious. I had invited Mary Jean Anderson to be my date at the upcoming junior prom, and she had accepted. This was to be our first date, and I desperately wanted it to be a success. Mary Jean was simply entrancing!

Anyway, because the custom at that time was for boys to buy their dates a corsage for all junior and senior proms, I headed off to a florist shop to purchase a corsage. I'd never bought a corsage before and was uneasy at the prospect. However, marshalling my courage—Mary Jean was clearly worth it—I entered the florist shop on the morning of prom night, where two middle-aged women were standing behind the counter. (At that time in my life, I regarded anyone older than 25 as middle-aged.)

I explained to the two clerks that I wanted to purchase a corsage for the prom. One of the women asked me how much I wanted to pay. I gave her a price range, happy that I had previously devoted some thought to that issue. Then the second clerk asked me what color my date's prom dress was. Again, having done plenty of advance thinking about this anxiety-laden floral encounter, I was prepared to respond. Actually, I had previously heard Mary Jane telling a friend about her prom dress. So, drawing on that knowledge,

I suavely replied, "She's going to wear an organdy dress." Hearing my response, both clerks frowned at me with undisguised disapproval. One of them said, "I'm sorry, young man, organdy is a fabric—not a color."

I had thought that organdy was a lavender-like color, but I subsequently learned it is a thin form of cotton used in curtains and formal evening dresses. My embarrassment was apparent to the clerks, doubtlessly due to the atypical color of my face which, I suspect, was bright organdy. I meekly said, "Can you please pick out a corsage?" They did. But my initial venture into the corsage-buying world had been stigmatized by a clear-cut case of gender bias.

What's biased about this item, I hope you already recognize, is that female executives in these fictitious companies are *always* paid substantially less than male executives. And, because the introductory remarks for this item fail to note the fundamental inequity of the salary data being presented, many students might be left to conclude that "This is perfectly normal and acceptable." If this item doesn't offend a fair number of female students, it certainly ought to!

Next, let's look at an item where girls will be unfairly penalized because they are less conversant than boys with a key concept in the item. Yes, once more we are dipping into the

world of YAC—the number of yards a football player gains after having caught a pass prior to being tackled. In Figure 4 you’ll find the kind of test item that might be employed to detect students’ abilities to interpret tables and graphs.

As you can see, there are two factors (YAC and age) that must be considered in order for a student to arrive at the best answer to this item. Many girls, who have no idea about the meaning of YAC, won’t know whether a large YAC average is good or bad. Moreover, girls are less apt to know that, by the time a professional football player gets to be in his mid-thirties, that player’s career usually is nearing its end. Thus, students who know about the longevity of professional football players, and who also know that YAC is not the sound made by a barnyard animal, will surely out-perform students on this item. And you guessed it: most of the students who will correctly identify that 24-year-old Billy Bob Boyd is destined for the highest salary will be boys.

Socioeconomic Bias, Assessment’s Closeted Skeleton

Description. *Socioeconomic bias* is far more widespread in educational assessment than is generally recognized, and it is

Figure 4. An Illustrative Data-Analysis Item Embodying Gender Assessment Bias

The National Football League has posted on its website the following YAC statistics for the league’s four leading receivers. The YAC averages, as well as the ages of these receivers, are presented below. Because all four players are free agents, each of their contracts will have to be renegotiated next year. Considering both the YAC data and the age data, select the receiver who is most likely to earn the highest salary next year. Circle the letter of your choice.

	YAC Average	Age
Jim Jones	15.4	35
Floyd Farmer	9.8	21
Willy Walker	12.7	29
Billy Bob Boyd	14.3	24

- A. Jim Jones C. Willy Walker
B. Floyd Farmer D. Billy Bob Boyd

particularly prevalent in large-scale exams such as nationally standardized achievement and aptitude tests. As is implied by its name, socioeconomic bias occurs when students are offended or unfairly penalized because of their family's socioeconomic status (SES).

Although it's always possible to offend low-SES students by including content in test items that denigrates impoverished parents, the vast majority of assessment bias that bubbles forth because of SES factors consists of unfair penalization. Children from less affluent backgrounds will find that, on certain test items, they don't have as much chance to succeed as their more affluent classmates. Because those diminished opportunities for success hinge not on a student's effort or ability but, rather, on the student's social or economic background, this clearly constitutes an instance of unfair penalization.

You may have noticed, two paragraphs earlier, I suggested that socioeconomic bias is frequently encountered in large-scale assessments such as nationally standardized achievement and aptitude tests. I'd better explain why that's so. Almost all nationally standardized educational tests are intended to fulfill a *comparative* measurement mission. That is, such tests are supposed to allow a test-taker's score to be contrasted with those of other test takers—typically the students who, having taken the test earlier, constitute the test's *norm group*.

Comparatively oriented tests permit us to determine that a given student scored high (for example, at the 96th percentile) or low (for example, at the 12th percentile) in relation to other test takers. But in order for these sorts of tests to provide their fine-grained comparisons, it is necessary for such tests to produce a considerable degree of *score-spread*—that is, a substantial scattering of students' total test scores. If students' scores are too tightly bunched together, comparative interpretations are difficult, if not impossible, to make.

Score-Spread: The degree to which test-takers' scores are dispersed

Okay, we see that score-spread is a necessary condition for large-scale tests to provide meaningful comparative interpretations. Well, one of the very best ways to make sure that achievement tests produce such score-spread is to include items on those tests that are linked to SES. SES-linked items are those on which children from more affluent families are likely to outperform children from less affluent families. Because SES is an educational variable that's spread out over a wide range (with plenty of low, middle, and high SES) and a variable that isn't altered all that rapidly, SES-linked items on large-scale tests do a terrific job of creating the score-spread these national tests require.

Do I think that the people who create large-scale assessments are members of a malevolent measurement mafia, an evil cabal intent on punishing low-SES youngsters? No, not at all. Instead, because these large-scale tests are often revised, sometimes many times, their SES-linked items tend to remain in new revisions of the tests. That's because SES-linked items do such a wonderful job of spreading out test-takers' scores. But SES-linked items, of course, constitute socioeconomic assessment bias, a form of bias flowing from factors over which a child has no control.

Illustrative items. Let's look now at a pair of test items suffering from serious cases of socioeconomic bias. Both of these illustrative items, incidentally, are based on items taken from currently used nationally standardized achievement tests. I've altered the items a bit, so I don't violate the security of the tests from which I took the items. But I assure you that the cognitive demands of the original items were identical to those contained in the two illustrative items you'll consider next.

Please take a gander at the illustrative item in Figure 5. As you can see, it's a science item and, in this instance, the item on which it is based was included in a nationally standardized test intended to measure sixth-grade students' science achievement. I'm sure you can see that the correct answer to this item is choice D, a telescope. But here's where SES spoils our assess-

Figure 5. An Illustrative Science Item Embodying Socioeconomic Assessment Bias

Suppose you wanted to determine if another planet had rivers or mountains on it. Which of the tools below would best help you find out?

- A. Camera
- B. Microscope
- C. Binoculars
- D. Telescope

ment stew: It's quite obvious that sixth-graders from affluent families are going to be more familiar with telescopes than are sixth-graders from impoverished families. Affluent families are more likely to have purchased a telescope for children. Affluent families are more likely to have cable television on which many science-related programs are available. Low-income families won't be able to afford telescopes or cable television. Kids from low-income families, therefore, are going to be unfairly penalized by an item such as the one seen in Figure 5.

As usual, think about assessment bias in terms of the way that large numbers of students would be likely to respond

to a test item. If 100 sixth-graders from affluent families and 100-sixth graders from low-income families completed the item in Figure 5, I'd bet big bucks that the affluent-family children would come out on top. Yes, some kids from rich families would answer the item incorrectly, and some kids from poor families would get the right answer. But the odds favor children from higher SES backgrounds. The item in Figure 5 assesses what children bring to school, not what they learn at school.

Incidentally, although the illustrative items I've used in this booklet are all of the selected-response variety (in this instance, a multiple-choice item), they could just as easily have been constructed-response items.

For instance, the science item in Figure 5 could have readily been transformed into a short-answer item in which the student would have been asked to *name* an appropriate scientific tool for spotting mountains or rivers on other planets. Such an item would, of course, still have been biased.

Now please consider the illustrative items in Figure 6. As you can see, it is a reading vocabulary item. The actual item on which I based this item is found in the reading

See also
*Selected-Response Tests:
Building and Bettering*

Figure 6. An Illustrative Reading Item Embodying Socioeconomic Assessment Bias

My mother's field is court reporting.

Choose the sentence below in which the word field means the same as it does in the boxed sentence above.

- A. The first baseman knew how to field his position.
- B. Farmer Jones added fertilizer to his field.
- C. What field will you enter after school is complete?
- D. The doctor checked my field of vision.

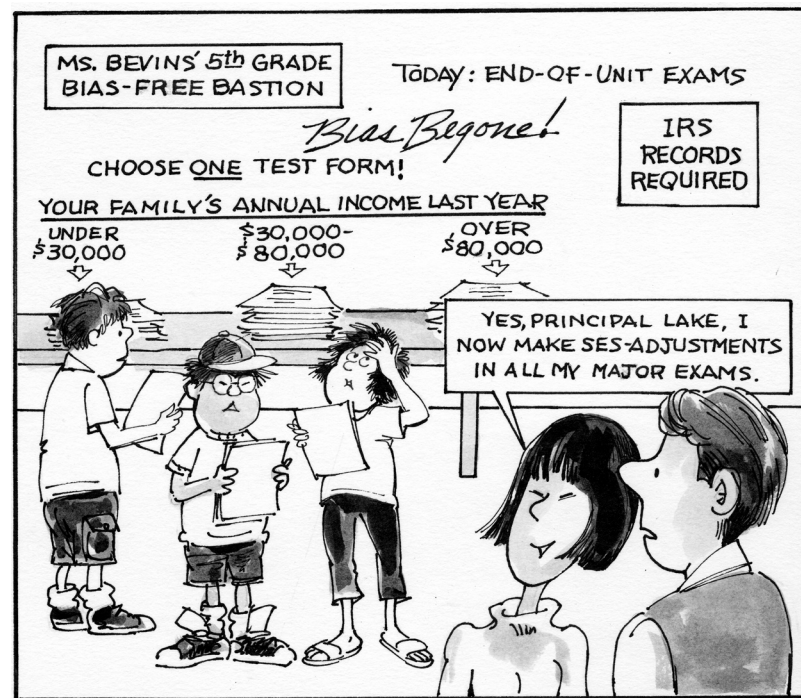
section of a fourth-grade nationally standardized language arts achievement test. The correct answer to the item, of course, is choice C, "What field will you enter after school is complete?"

Please think for a bit about 100 fourth-graders from low-income families and 100 fourth-graders from high-income families. Isn't it more likely that the children from the high-income families will have one or more parents whose employment is usually described as a *field*. I'm thinking of doctors, lawyers, computer programmers, teachers,

and so on. On the other hand, many of the parents from low-income families will have *jobs*, not fields. If you're a child whose mother works as a cashier in a mini-mart, and whose father works in a car-wash, your parents don't have fields. Can you see, then, that low-SES kids would be unfairly penalized by this sort of test item?

Disturbingly, you'll find many more of these sorts of items on nationally standardized tests than you'd expect. And even on state-developed or district-developed achievement tests—tests never intended to yield fine-grained comparative interpretations—you'll often discover that inexperienced measurement specialists have sometimes allowed SES-linked items to flourish in a misguided effort to promote the sort of technically applauded score-spread that's found in national tests.

To wrap up this quick look at three common sources of assessment bias, I need to remind you that bias can surely arise—and often does—from sources other than racial/ethnic, gender, and socioeconomic factors. For instance, assessment bias can spring from children's religions, the geographic locales in which they live, and even their parents' political or religious preferences. Bias seems to lurk in every assessment shadow, eager to trip up anyone who tries to derive accurate test-based inferences about students.



BIAS DETECTION

I'd like to deal, very briefly, with one final topic in this booklet. It concerns how educators can go about identifying and eliminating assessment bias, no matter what that source of bias is. Essentially, the detection of bias in educational tests boils down to one of two approaches: *judgmental* and *empirical*. Let's take a quick look at each.

Judgmental Approaches

When high-stakes examinations are developed these days, there's invariably an item-by-item scrutiny of the test's under-development items (that is, still in their draft versions) by a group of individuals usually described as a "Bias Review Panel" (or something similar). Typically, this panel is composed of educators and noneducators who represent the largest minority student subgroups who will, in the future, be required to take the test for which the items are being bias-reviewed. There should also be a reasonable representation of males and females on such panels.

Typically, the panel's members, after having received an orientation and some training, will review every potentially usable item by supplying a *Yes* or *No* answer to a review question such as the following:

A Typical Per-Item Judgment Question for Members of a Bias Review Panel

Might this item offend or unfairly penalize any group of students on the basis of such personal characteristics as race, gender, ethnicity, or religion? (Please answer Yes or No for each item.)

Often, reviewers are asked not only to render a *Yes/No* judgment but also to supply a written comment noting the nature of any bias the panelist believed was present. This "nature of the bias" comment can help identify irrelevant judgments such as "This item is too difficult for anyone."

I'd like you to look at the wording in the preceding item-judgment question so you can see how a slight alteration in the language of such an item-review question is likely to influence the nature of reviewers' judgments. Note that the question commences with the word *might*. If the question had, instead, led off with the word *would*, then think about the influence of that one-word alteration on reviewers' judgments. Which of those two questions—the "might" question or the "would" question—will most likely lead to more items being identified as biased by reviewers?

Well, I've actually used both forms of such a question with numerous bias review committees, and I can assure you that the "might" version definitely results in more bias-is-present judgments from reviewers. You see, when there's a "would" in the question, this requires a definitive level-of-certainty judgment from reviewers that's clearly more difficult to arrive at than when there's a "might" in the review question. The "might" version of the question says to reviewers that if there's even a *possibility* of assessment

bias in the item, then the reviewer needs to judge the item adversely.

Usually, when item review panels scrutinize items, the panelists make their judgments independently for each item. Then, later, the reviewers' independent judgments are summarized so that any items receiving a predetermined minimum number of "bias present" judgments will be excised from the pool of eligible items. Often, if there is great concern about bias removal, any item that received *even one* adverse judgment is rescutinized, sometimes by a smaller committee of bias reviewers, to make sure that no biased item has escaped detection.

For bias review of classroom tests—that is, those constructed by teachers themselves—the key elements of the judgmental appraisal used by bias review panelists is applicable. Teachers should judge each of their items on the basis of whether an external reviewer would, for that item, supply a *Yes* or *No* judgment to the kind of review question given a few paragraphs earlier.

It is also beneficial for teachers to call on a colleague of a particular minority group to help them identify what warrants special attention when trying to ferret out racial/ethnic assessment bias for that particular subgroup of minority students. It is unrealistic to think that a teacher will be sensitive to every potential bias. Constructive colleagues can be a big help.

Empirical Approaches

These days, the most common empirical approach to bias detection consists of analyzing test items based on their use in operational test forms or, perhaps, based on their administration, as part of a limited field test in advance of the test's official use. These analyses try to identify *differential item functioning* (DIF), and several versions of DIF analyses are currently in common use—all of which rely on sophisticated computer machinations. The thrust of DIF analyses is to see whether any differences in the performances of subgroups on each item are substantial enough to exceed what might be expected on the basis of mere statistical chance.





Actually, DIF is just a far more sophisticated way of figuring out whether a disparity between student groups' success on a test item is sufficiently substantial that, at least on empirical grounds, the item is identified for further scrutiny (perhaps by a bias review panel). In the old days, when we were just beginning to use students' item performances to decide if an item might be biased, all we would do is calculate a series of *p*-values for the item (the percent of students who answered a particular item correctly). We'd determine a *p*-value for African American students and a *p*-value for white students, then see if there was a difference between the *p*-values and,

if so, in which direction and how large. Trivial p -value disparities were usually ignored. However, if on a given item the p -value for white students was .79, and the p -value for African American students was only .37, then that p -value difference of .42 sent up all sorts of flags. These days, DIF analysis accomplishes the same mission, but more accurately because it takes into consideration not only a subgroup's performance on a single item but also that same subgroup's performances on all the other items in the test. DIF analyses, as a consequence, are far more precise than their p -value-based predecessors.

Unfortunately, all empirical analyses of potential test bias require the use of fairly large student samples (in order to attain sufficient statistical accuracy). Classroom teachers almost never have enough students representing particular subgroups so that empirically based bias-detection techniques can be used. Therefore, teachers are really left with only judgmental approaches to use in their bias-detection efforts. But, as I suggested earlier, the very best way for a teacher to tell if an item offends or unfairly penalizes students on the basis of personal characteristics is for the teacher to be *extraordinarily attentive* to the possibility that assessment bias has somehow set up a base-camp in one of your tests.

RECAP AND WRAP-UP

Back at the beginning of this booklet, I confessed in advance that I wanted you to understand:

-  ***What constitutes assessment bias and the two ways such bias can reduce the validity of test-based inferences.***
-  ***That “disparate impact” of an educational test does not automatically signify the presence of assessment bias.***
-  ***The nature of three common sources of assessment bias: racial/ethnic bias, gender bias, and socioeconomic bias.***
-  ***How assessment bias can be reduced in both large-scale tests and classroom tests.***

Because assessment bias tends to squeeze the validity out of test-based inferences about students, assessment bias should clearly be expunged from educational tests. But expunging is easier to advocate than it is to accomplish. Even so, the first two steps in reducing

assessment bias are (1) to recognize its prevalent and insidious existence and (2) to realize it will take more than rhetoric to get rid of it. This booklet, in a sense, attempts to contribute to both of those significant steps.

Assessment bias was defined as the content of any test item that offends and/or unfairly penalizes students on the basis of such personal characteristics as students’ race, gender, or religion. Three commonly encountered forms of assessment bias—racial/ethnic bias, gender bias, and socioeconomic bias—were described and illustrated.

Disparate impact occurs when a test item leads to unequal performances from different groups of students, such as is seen when there are large differences between the success rates of boys and girls on a particular item. It was stressed that disparate impact, all by itself, does not indicate assessment bias. Rather, the existence of disparate impact in an item strongly indicates that the item should then be carefully judged to determine if the item is biased or if there are shortcomings in the instruction previously provided to a low-performing subgroup of students.

Finally, two approaches to detecting assessment bias were briefly considered: judgmental and empirical bias-detection strategies. For classroom use, because there is usually an insufficient number of students constituting the subgroups involved, only judgmental review procedures are routinely applicable.

The chief message in this booklet about assessment bias is that teachers must become extraordinarily alert to the possibility that assessment bias has crept into their teacher-made tests. Because assessment bias offends or unfairly penalizes students, and thus reduces the validity of test-based inferences about those students, we need to eliminate as much assessment bias as we possibly can.

GLOSSARY TERMS

Assessment Bias: The presence of one or more items in a test that offend or unfairly penalize students because of students' personal characteristics

Disparate Impact: When an item on a test results in substantial differences in the success rates of separate subgroups of students

Offensiveness: Test items containing content that insults, irritates, or causes pain to students because of those students' personal characteristics

Score-Spread: The degree to which test-takers' scores are dispersed

Unfair Penalization: Test items containing content that unjustly prevents one or more subgroups of students from performing well because of those students' personal characteristics

REFERENCES

Jamal Abedi. “The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues.”

Educational Researcher, 33, no. 1 (January/February 2004): 4–14.

Paul E. Barton. “Why Does the Gap Persist?”

Educational Leadership, 62, no. 3 (November 2004): 8–13.

Kenneth Carlson. “Test Scores by Race and Ethnicity.”

Phi Delta Kappan, 85, no. 5 (January 2004): 379–380.