ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/265086461

Test fairness

Chapter · January 2006

CITATIONS
71

READS
7,967

1 author:

Gregory Camilli
Law School Admission Council
96 PUBLICATIONS **3,720** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Understanding Influences on Mathematics Achievement View project

Project Law School Studies View project

All content following this page was uploaded by Gregory Camilli on 19 July 2016.

The user has requested enhancement of the downloaded file.

7

Test Fairness

Gregory Camilli
Rutgers, The State University of New Jersey

1. INTRODUCTION

This chapter is about the social context of test fairness, and methods of logical and quantitative analysis for examining test fairness. While there are many aspects of fair assessment, it is generally agreed that tests should be thoughtfully developed and that the conditions of testing should be reasonable and equitable for all students. Many of the chapters in this volume address substantive issues related to this concern; and in particular, the chapter by Kane on validity, and reliability, in the chapter by Haertel on reliability, describe concepts and tools that can be understood as prerequisites for test fairness. Beyond this foundational material, issues of fairness involve specific techniques of analysis, and the more common approaches are reviewed in this chapter.

Some important concepts of social and legal and social justice are summarized initially to provide both context for modern concepts of fairness, and insight into the logic of analyzing fairness. This is important because many unfair test conditions may not have a clear statistical signature, for example, a test may include items that are offensive or culturally insensitive to some examinees. Quantitative analyses infrequently detect such items; rather, such analyses tend to focus on the narrower issues of whether a measurement or prediction model is the same for two or more groups of examinees. Indeed, these statistical models can be the same even when a test or assessment has substantial negative consequences for the members of some groups of test takers. This ostensible contradiction might result from a technically sound test that is developed with a carefully delimited set of purposes, yet used in a manner that is inconsistent with those purposes.

An exhaustive treatment of issues related to fairness is beyond the scope of this, or any single chapter. Rather, I have provided a conceptual framework for understanding both test fairness and the assumptions and modes of inferences that underlie corresponding statistical analyses. Recent methodological developments are examined as well as item sensitivity review, fairness in classroom assessment, and historical themes regarding fairness in college admissions. The current chapter greatly benefited from a number of sources including the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999); the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988); the *Code of Professional Responsibilities in Educational Measurement* (NCME Ad Hoc Committee on the Development of a Code of Ethics, 1995); *Educational Measurement, Third Edition* (Linn, 1989); and the work of Hartigan and Wigdor (1989), Holland and Wainer (1993), Hubin (1988), and Camilli and Shepard (1994).

Most of the above sources deal largely with test fairness in the United States, though test fairness is obviously not just an American concern. Fairness issues are inevitably shaped by the particular social context in which they are embedded. Thus, while the psychometric methodologies reviewed herein generalize broadly, the substantive generalizations may vary with respect to other cultures and countries. One special emphasis herein is on issues of fairness with respect to race and ethnicity. Many other topics are not considered due to space limitations but are clearly as important including: gender; special populations; licensing and certification; test accommodations; and linguistic diversity. While some of these are treated indirectly, many of the principles in this chapter are directly applicable. Finally, the issue of fairness in high-stakes testing in public K-12 education is not examined. This is a topic unto itself, requiring more space than that of a single section within a broad chapter.

2. SOCIAL AND LEGAL CONTEXT OF FAIRNESS ISSUES

Two fundamental premises of a liberal society are that a republic is a voluntary association between free individuals, and that a free-market economy is based upon fair competition. *Liberal*, in the original sense of the word, referred to the antithesis of social systems based on hierarchy and subordination. In a liberal society, free individuals are members, not subjects, of the state, and government is constituted by the members of society for their mutual benefit. This precept is embodied in the foundational text of the United States, the Declaration of Independence, which holds that "that all men are created equal, that they are endowed by

611-190-007.indd 221 7/14/2006 4:18:15 PM

their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness." And, importantly, "That to secure these rights, Governments are instituted among Men, deriving their just Powers from the consent of the governed. . . ."

In a liberal society, the power of the federal government in regulating social transactions is distributed by design. This idea is embodied in U.S. Congressional representation, which defines "equal representation," in terms of both individual citizens (through the House) and states (through the Senate). However, individual rights were first addressed in the Bill of Rights (the first 10 amendments to the Constitution), which restricted contravention by federal government of individual liberties, including those of freedom of speech, press, assembly, religion, and due process for those accused of committing a crime. Modern interpretations of equality are more directly traceable to the 13th and 14th Amendments, which were designed to dismantle the institution of slavery.

Equality, in the senses of *fair play* and fair competition, was defined with the *individual* as the fundamental unit of both government and society. Ideally, this would have resulted in a *meritocracy* (originally a pejorative term coined in 1958 by Michael Young in the book *The Rise of Meritocracy*), but the scope of individual liberty was limited in early America. For example, women could not vote until 1920, and though all African Americans were legally freed in 1866, another 100 years passed before the Supreme Court decision *Brown v. Board of Education* (1954) began to unravel the apartheid system of racial segregation as established by laws of the Jim Crow era. As opposed to monarchical societies, however, early American political philosophy espoused the principle of accrual of the benefits of society based on merit. Fair competition allows individuals to seek happiness as they prefer *and* allows for the actual accrual of benefits to be unequal. It is only rules and standards for competition that should be impartial to all individuals.

Though some form of merit is frequently appropriate in attaining the benefits of a society (though not all, consider healthcare), there are pitfalls. Young (2001) argued that whereas historically ability was more-or-less randomly distributed among social classes, such ability might become highly concentrated within particular social "castes" if merit is narrowly defined. This may result in potential leaders, artists, scientists, scholars, or technicians being distanced from their various cultural and regional affiliations. Hartigan and Wigdor (1989) recognized alternate conceptions of fair play that characterize equality as more than fair competition, or alternatively, the absence of irrelevant barriers. Instead, they argued that equality may be fostered when individuals with "similar talents have similar life chances" and

> A much more radical interpretation of equal opportunity might call for equalizing the conditions of the development of talent throughout society so that all children enjoy the same material and cultural advantages. (p. 33)

To some, but not others, fairness requires a vision beyond that of competition as embodied by legal statutes of due process. The idea of unqualified individualism (Jensen, 1980), from this perspective, lacks the moral impetus for promoting social change.

2.1. The Fourteenth Amendment

Just after the Civil War, in 1868 the "Radical Republican" members of the 39th Congress intent on protecting newly emancipated African Americans, championed establishment of the 14th Amendment for insuring equal justice under the law. The relevant first section reads,

> All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States and of the State wherein they reside. No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.

Congress, aware of potential resistance in ratifying the 14th Amendment, provided that both state congressional representation and tax allocations would be proportionally decreased according to the number of citizens (excluding Native Americans and, implicitly, women) denied the right to vote. Resistant state governments were eventually compelled to ratify the 14th Amendment, though tangible acceptance of the legal requirements was difficult and slow. This notion of *equal protection* (a close relative of *fair play*) became critical again in the 1950s, 1960s and 1970s for challenging segregation practices ranging from voting rights to school attendance to public accommodations to employment practices. According to Hartigan and Wigdor (1989), this amendment has been the major vehicle for developing substantive meaning for the concept of equality in our society.

Constitutional amendments, as noted above, were limited to infringements by state governments (and their subordinate entities, such as cities, counties, and school boards) and did not pertain to actions by nonstate and private sector entities. The next milestone in legal deliberations on the concept of fairness was the Civil Rights Act of 1964, which was intended

> To enforce the constitutional right to vote, to confer jurisdiction upon the district courts of the United States to provide injunctive relief against discrimination in public accommodations, to authorize the Attorney General to institute suits to protect constitutional rights in public facilities and public education, to extend the Commission on Civil Rights, to prevent discrimination in federally assisted programs, to establish a Commission on Equal Employment Opportunity, and for other purposes.

This law, which was the most significant civil rights legislation since reconstruction, went beyond Congressional enforcement of the 14th Amendment against discriminatory employment actions by state governments. Congress, using its power to regulate interstate commerce, prohibited discrimination based on "race, color, religion, sex or national origin" in public establishments. In the employment context, the Civil Rights Act (Title VI) is enforced by the

611-190-007.indd 222 7/14/2006 4:18:17 PM

Equal Employment Opportunity Commission (EEOC). In the public schools context, the law (Title VI) is enforced by the Office of Civil Rights (OCR). Similarly, Title IX of the Education Amendments of 1972 (20 USC §§ 1681–1688), prohibits discrimination on the basis of sex in education programs and is enforced by the OCR. Most of these laws apply to programs and activities that receive federal financial assistance, while the 14th Amendment applies notwithstanding receipt of federal funding.

2.2. Adverse or Disparate Impact

Title VII of the Civil Rights Act of 1964 concerns discrimination in employment practices, and the Equal Employment Opportunity Commission (EEOC) was created to provide leadership and enforcement regarding Title VII. Section 703(a)(2) declares it unlawful

> to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.

According to Hartigan and Wigdor (1989), the nascent EEOC in 1966 interpreted Title VII discrimination to consist of employment practices *intended* to discriminate or to treat people of protected status differently from others, but includes as well practices having a harmful affect on members of protected groups.

Title VII legitimates challenges to employment practices that adversely affect such groups—whether the outcome was intentional or not (*Griggs v. Duke Power Company*, 1971). *Adverse impact*, which is a measurable outcome related to the Title VII term "adversely affect," has come to denote the selected proportions of people from groups with protected status relative to the largest selection proportion. Yet Congress did not require employers to correct imbalances in their workforce that could not be demonstrated to result from discriminatory practices. In fact, Section 703(j) states specifically that no requirement exists for granting preferential treatment to any individual or group based on extant imbalance with regard to race, color, religion, sex, or national origin.

Adverse impact in employment decisions has been defined as a substantially different rate of selection that creates an imbalanced workforce with respect to a group with protected status. (Note that the phrase "disparate impact" is used in the Civil Rights Act of 1991, PL 102-166). In particular, the usual threshold for adverse impact is established if the selection rate for one group is less than 80% (known as the 4/5s rule) of that for the group with the highest selection rate (EEOC's Uniform Guidelines on Employee Selection Criteria). However, courts have not rigidly interpreted statistical criteria for demonstrating adverse impact, and adverse impact by itself is not a sufficient basis for establishing violation of equal protection. Practices that result in adverse impact are deemed unlawful only (a) if the practice cannot be demonstrated to be job-related and "consistent with business necessity" or (b) if an alternative employment practice with less disparate impact exists and the "respondent refuses to adopt such alternative employment practice." The demonstration of disparate impact establishes the grounds for legal challenge, that is, a *prima facie* argument, and the effect of Title VII was to place the burden of proof on employers in the context of adverse impact given a prima facie case (e.g., one demonstrating the 4/5s rule).

It is important to recognize that the applicability of this law was reduced in 2001 when the Supreme Court ruled in *Alexander v. Sandoval* (2001) that disparate impact arguments could not be brought to federal courts by individual citizens, as had been the case for the previous 35 years. This has effectively rendered disparate impact viable only in the context of OCR administrative enforcement actions (Welner, 2001), and this opinion may reduce the number of venues in which psychometric evidence of disparate impact is salient. It has, nonetheless, no bearing on professional standards for fairness (reviewed in section 3.2). Rather, this course of events serves to illustrate that legal standards and professional responsibilities for ensuring fairness are not necessarily commensurate.

2.3. Individuals Versus Groups

Titles VI and VII represented a dramatic and controversial move beyond the 14th Amendment. As recognized by Hartigan and Wigdor (1989),

> A persistent anomaly in federal civil rights policy has been the adherence, on one hand, to the principle that the Constitution and Title VII protect the rights of *individuals*, and the adoption, on the other, of a definition of discrimination that looks to the effects of employment procedures on *groups*. (pp. 39–40)

Title VII was intended to protect individuals rather than groups, but requires evidence based on the classification of individuals into protected groups. One resulting irony is that discrimination might be demonstrable statistically at the group level, but not in the case of any particular individual. Under Title VII, which is a statutory law, "group" evidence is admissible even though the very same evidence would not be admissible under constitutional law. Thus, statutory law modified the original notion of individual due process in the 14th Amendment: for some purposes, a person on the playing field could now be classified, rather than evaluated individually.

As of 1997, the federal government recognized a number of official group classifications for federal reporting purposes:

2.3.1. Designation of Race

In October 1997, the Office of Management and Budget (OMB) released new categories for collecting data on race and ethnicity (OMB, 1997). The new racial categories (see Table 7.1) established were White; Black or African American; Asian; Native Hawaiian or Other Pacific Islander; and American Indian or Alaska Native. In contrast to *racial* categories, several *ethnic* categories were designated. For designating ethnicity, the OMB categories are Hispanic or Latino and Not Hispanic or Latino. While OMB did not

TABLE 7.1 OMB Guidelines for Federal Reporting on Race and Ethnicity (see OMB, 1997)

Race or Ethnicity	Description
American Indian or Alaska Native	A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
Asian	A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
Black or African American	A person having origins in any of the black racial groups of Africa. Terms such as "Haitian" or "Negro" can be used in addition to "Black or African American."
Hispanic or Latino	A person of Cuban, Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race. The term, "Spanish origin," can be used in addition to "Hispanic or Latino."
Native Hawaiian or Other Pacific Islander	A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
White	A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

offer a "multiracial" category, it allowed that individuals could select one or more races—recognizing that designations of race are becoming more complex and nuanced.

2.3.2. Designation of Ethnicity

One general definition is that an ethnicity is a group of people who see themselves as sharing one or more of the following: language, race, place of origin, values, and history. Members of such a grouping might also differentiate themselves from other groups based on these characteristics. A shorter definition of an ethnicity is a group that shares or is perceived to share a culture. In this sense, "culture" broadly defined is a common heritage or set of beliefs, norms, meanings, and values. A variety of ways exists to define a cultural grouping (e.g., by ethnicity, religion, geographic region, age group, sexual orientation, or profession), and many people consider themselves as having multiple cultural identities. Increasingly, self-identification appears to be driven by factors other than the traditional concept of race.

According to the U.S. Census, there are many distinct ethnicities, and examples may be more useful than definitions in conveying the meaning. Hispanic American (Latino) is an ethnicity according to OMB and may apply to a person of any race, and many people from Caribbean nations identify their ethnicity as Hispanic or Latino and their race as African American. Within the broad category of "Asian Americans and Pacific Islanders," there are 43 ethnic groups speaking over 100 languages and dialects. For "American Indians and Alaska Natives," the Bureau of Indian Affairs currently recognizes 561 tribes.

2.3.3. Other Federal Categories

According to Evinger (1995), the OMB categories were developed largely for collecting data from groups in the United States that have historically suffered discrimination and adverse treatment. Rather than implying a static racial classification, categories are assumed to exist relative to disparate experience.

Under Federal OMB reporting requirements (OMB, 1997), neither gender nor limited English proficiency (LEP) is an official designation. Other laws and regulations (e.g., No Child Left Behind, PL 107-110) require desegregation of data by categories such as grade or gender. The term "protected group" can apply to all of the official designations, and may incorporate other groups established on the bases of equity concerns. For example, an analysis of fairness could designate a protected group as children from highly mobile families, children with special education needs, or even students who are frequently absent.

3. THE STANDARDS FOR TEST FAIRNESS

The 1999 *Standards for Educational and Psychological Testing* (*Standards* for short) followed two previous editions in 1974 and 1985. Revision of the 1985 *Standards* was undertaken by a joint committee of the American Educational Research Association, American Psychological, and National Council for Measurement in Education, and was the first edition to contain explicitly a chapter on test fairness. Comments on draft versions of the 1999 S*tandards* were received from 72 organizations including professional societies, credentialing boards, government agencies, test developers, and academic institutions. As noted in the *Standards*, its purpose is to provide criteria for the evaluation of tests, testing practices, and the effects of test use, where the term *test* here is used to specify a broad range of standardized assessments including particular tests, scales, inventories, and instruments. The term *standardized* here refers not to multiple-choice items, as many misunderstand, but to conditions of test administration that are equivalent for all examinees.

The 1999 *Standards* is a document that reflects a commitment by measurement professionals to address social as well as technical concerns of fairness. This document reflects a broad, rather than a narrow vision:

> A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society. It would consider the technical properties of tests, the ways test results are reported, and the factors that are validly or erroneously thought to account for patterns of test performance for *groups and individuals.* [emphasis added] (p. 73)

Tests designed properly and used fairly can facilitate positive educational, social, and economic goals. However, tests are only one source of evidence for informing decisions. For example, in merit-based selections, tests scores are known to correlate modestly with criterion measures of success (e.g., first-year college GPA), and the measurement community has cautioned that test scores are inadequate as a surrogate for merit (Wightman, 2003).

3.1. Relevant Definitions

Prior to considering fairness issues, a number of terms are introduced below including: fairness, statistical bias, disparate impact, fairness, differential item functioning, and test bias. Whereas most of these terms have relatively specific meanings, the term fairness has a much broader usage:

> *Fairness* in testing refers to perspectives on the ways that scores from tests or items are interpreted in the process of evaluating test takers for a selection or classification decision. Fairness in testing is closely related to test validity, and the evaluation of fairness requires a broad range of evidence that includes empirical data, but may also involve legal, ethical, political, philosophical, and economic reasoning.

As noted in the 1999 *Standards*, fairness "is subject to different definitions and interpretations in different social and political circumstances" (p. 80). There are different ways to describe fairness in the context of predictive models for selection (section 5.2), but there are also different ways to conceptualize fairness *per se*.

Given the ubiquity of "groups" in establishing fairness, several definitions are required for the situation in which two groups are compared. Following Holland and Thayer (1988, p. 130),

> The performance of two groups may be compared in terms of a test item, a total test score, or a prediction regarding success on a criterion. The *focal* group (or group F), which is sometimes called the *protected* group, is of primary interest. This group is to be compared to a second group, labeled the *reference* or *base* group (or group R). The previous terminology of minority and majority groups is no longer used.

In typical fairness analysis, the first step is to define groups and then to compute differences between groups R and F in terms of percent selected (say for a job or college position), or average scores on an item or on a test. The obtained difference is described with the label disparate impact:

> In legal analysis, *disparate impact* describes group differences in test performance that result in different group proportions of candidates identified for selection or placement. In studies of test fairness, the term *impact* has been borrowed, and is used to describe the observed difference between the average scores for two groups on a particular test or test item.

Once group differences are estimated, the important task is to distinguish a genuine group difference in proficiency from one that arises from a distorted measurement process. An observed difference (i.e., impact) does not necessarily imply measurement bias.

Disparate impact, to give a simple example, can be thought of as the difference in average running speeds over a fixed distance for two groups using an accurate stopwatch. This difference signals that one group on average is faster than the other. In contrast, statistical bias arises when two groups (R and F) are in reality highly similar in average running time, but stopwatch A for group R is accurate, while the stopwatch B for group F is inaccurate. In the latter case, the running speeds for individuals in group F will be "biased." If it is further established that stopwatch B runs too fast for group F, then the timing information provided by the stopwatch B is *unfair* in the sense that group F is disadvantaged. If stopwatch B runs too slowly, the timing information is *unfair* in the sense that group F is advantaged. In any event, the rankings of runners *within* two groups may be accurate, but comparisons between the best runners from each group and comparisons between group averages would be confounded with the inaccuracy of stopwatch B.

In this chapter, two terms are distinguished that are often used interchangeably elsewhere: fair and unbiased. I use the term *bias* synonymously with the phrase *statistical bias*, except as noted. Specifically,

> *Statistical bias* refers to the under- or over-estimation of one or more parameters in two kinds of statistical models—
>
> 1. In the field of measurement, parameters generally describe properties of examinees or items. Examinee parameters are referred to as person parameters or proficiencies. Item parameters come in three varieties: difficulty, discrimination, and pseudo-guessing.
> 2. In predictive models, parameters describe the intercepts and slopes of regression equations as well as the error variance.

The word bias in this context is reserved for a purely formal concept. This may seem overly technical, but it is better to consider bias as a measurement anomaly, and then to consider fairness in terms of who is advantaged or disadvantaged by the anomaly. Broadly speaking, statistical bias can be thought of as a systematic difference between two parameters that should be equal. Though all estimates embody some level of *random error*, bias is a kind of *systematic error*.

Disparate impact implies neither statistical bias nor unfairness. Differential group performance may be attributable to *bona fide* differences between groups in terms of the test construct. Test fairness does not imply equal outcomes. The purpose of a fairness investigation is to sort out whether the reasons for group differences are due to factors beyond the scope of the test (such as opportunity to learn) or artifactual. For example, immigrant children often score lower than indigenous students do on math items requiring

611-190-007.indd 225 7/14/2006 4:18:18 PM

more proficiency in the dominant language. The conclusion that the gap in performance on such an item signified bias would be warranted only if it could be demonstrated that immigrant children and indigenous children of *comparable mathematics proficiency* performed differentially on the item). This leads directly to the definition of statistical bias with respect to item performance:

> *Differential item functioning* or DIF, for short, is said to occur when examinees from groups R and F have the same degree of proficiency in a certain domain, but difference rates of success on an item. The DIF may be related to group differences in knowledge of or experience with some other topic beside the one of interest.

An item that does not show DIF has equivalent measurement properties for groups R and F, or, alternatively, the item is *measurement invariant*. In contrast, DIF occurs when an item's difficulty parameter is different for groups of comparable proficiency. Note that the definition above is also an attempt to shed light on why DIF occurs, a topic is given more attention in section 4, and sections 6 through 9.

The last definition is for the situation in which examinees are selected with a qualifying examination, and the qualifying score is used to make a prediction regarding the candidates' likelihood of success on a criterion. For example, LSAT (Law School Admission Test) is used to predict first year grade point average (GPA) in law school. Then, as Cleary (1968, p. 115) suggested,

> *Test bias* occurs for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance.

Parallel to DIF, test bias is also referred to as *differential prediction of a criterion*, or simply differential prediction. Though this definition is widely accepted among psychologists, the term "test bias" is often less descriptive of fairness concerns than the term "selection bias," a point addressed in more detail below. Differential prediction results from bias in measurement or regression parameters; for example, a regression prediction might be two high or too low if the intercepts or slopes of two groups differ and a common regression equation is used. This topic is more fully explored in sections 4 and 5.

3.2. Test Fairness Standards

In this section, an overview is provided of the twelve criteria pertaining to fairness given in the 1999 *Standards for Educational and Psychological Testing*. These appear in Table 7.2.

We note that of the twelve criteria, nine refer to groups or subgroups and three to individuals or aspects of individuals. Along a second dimension, six standards refer to interpretation, reporting, or use of test scores; three to *differential measurement or prediction*; and three to equity or sensitivity. Recall that differential measurement and differential prediction are consistent with the statistician's use of the term *bias*.

The standards regarding test fairness are classified in Table 7.3 according to this heuristic scheme. Standard 7.1 appears in two cells because it is concerned with interpretation and use if differential measurement is detected. Hard-to-classify standards include 7.10 because it is concerned with responsibility for collecting evidence of construct representation in the presence of group differences, and 7.7 because it concerns language, but not different language groups. Standard 7.3 covers *differential item functioning*, and Standard 7.6 covers *differential prediction*. These two aspects of statistical bias are considered in separate sections below. The upper middle cell in Table 7.3 contains standards regarding score interpretation and used, and appropriate limitations on test score use when statistical bias is observed.

Individual fairness requires standardized conditions of testing in which students are treated comparably. This type of fairness is denoted as *equity* (as in Standard 7.12). Standards 7.5 and 7.7 are variations on this theme with respect to language (or reading) and score interpretation. Denotation of group membership is neither required nor implied. The other standards explicitly appeal to group membership—but are not limited to OMB categories (see section 2.3). Evidence can be collected for alternative group definitions, depending on logical assessments of potential adverse or disparate impact. In addition to race and ethnicity, groups (depending on one's purpose) may be identified by social class, age, regions, urbanicity, and so forth.

If a test item is equitable, it is presented to individuals under impartial conditions, meaning that no student is favored over another in answering the item. If a test item is fair, it is (a) invariant across groups with respect to measurement and prediction *and* (b) equivalent across groups with respect to presentation, interpretation, reporting at the item level, and summative or formative use. For both test scores and test item responses, an additional concern regards racial, cultural, and ethnic sensitivity. Testing conditions and test content should avoid stereotyping, culturally offensive material, and other negative implications. Sensitivity problems may lead to statistical bias and thus faulty interpretation of test scores, but they can also be damaging or hurtful to individuals taking tests. The process of screening items in this regard is called *sensitivity review* (see section 10).

3.3. Test Validity and Social Constructions

The 1999 *Standards* stress that validity is addressed by collecting a variety of evidence. Likewise, claims that a test is fair must be supported with evidence. In addition to the general principles of validity and reliability, this requires evidence that the conditions of testing are equitable, and that scores on a test have the same meaning for different groups of examinees. An additional category of evidence concerns the consequences of testing, yet in this regard,

611-190-007.indd 226 7/14/2006 4:18:18 PM

TABLE 7.2 Fairness Standards from the 1999 *Standards for Educational and Psychological Testing*

7.1	When credible research reports that test scores differ in meaning across examinees subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup. Subgroups may be found to differ with respect to appropriateness of test content, internal structure of test responses, the relation of test scores to other variables, or the response processes employed by individual examinees. Any such findings should receive due consideration in the interpretation and use of scores as well as in subsequent test revisions.
7.2	When credible research reports differences in the effects of construct-irrelevant variance across subgroups of test takers on performance on some part of the test, the test should be used if at all only for those subgroups for which evidence indicates that valid inferences can be drawn from the test scores.
7.3	When credible research reports differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, the test developers should conduct appropriate studies when feasible. Such research should seek to direct and eliminate aspects of test design, content, and format that might bias test scores for particular groups.
7.4	Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the construct.
7.5	In testing situations involving individualized interpretations of test scores other than selection, a test taker's score should not be accepted as a reflection of standing on the characteristic being assessed without consideration of alternative explanations for the test takers performance on that test at that time.
7.6	When empirical studies of differential prediction of a criterion for members of different subgroups are conducted, they should include regression equations (or an appropriate equivalent) computed separately for each group or treatment under consideration or an analysis in which the group or treatment variables are entered as moderator variables.
7.7	In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct.
7.8	When score are disaggregated and publicly reported for groups identified by characteristics such as gender, ethnicity, age, language proficiency, or disability, cautionary statements should be included whenever credible research reports that test scores may not have comparable meaning across these different groups.
7.9	When tests or assessments are proposed for use as instruments of social, educational, or public policy, test developers or uses should proposing the test should fully and accurately inform policymakers of the characteristics of the tests as well as any relevant and credible information that may be available concerning the likely consequences of test use.
7.10	When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test scores difference between relevant subgroups of examinees should, where feasible, be examined for subgroups for which credible research reports mean difference for similar tests. Where mean differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct misrepresentation of construct-irrelevant variance. While initially the responsibility of the test developer, the test users bears the responsibility for uses with groups other than those specified by the test developer.
7.11	When a construct can be measured in different ways that are approximately equal, in their degree of construct representation and freedom for construct-irrelevant variance, evidence of mean score difference across relevant subgroups of examinees should be considered in deciding which test to use,
7.12	The testing or assessment process should be carried out so that test takers receive comparable and equitable treatment during all phases of the testing or assessment process.

there is currently no consensus on evidentiary requirements among measurement professionals. Both validity and reliability are considered in more detail in Chapters 2 and 3. Below, a brief treatment of construct definition and testing consequences serves as a backdrop for understanding how the choice of test construct and interpretation of test scores interact with fairness issues.

3.3.1. Construct Definition and Terminology

Test development begins with specifying a *construct*, which can be defined as "the concept or characteristic that a test is designed to measure" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The notion of construct validity is then derived as the

TABLE 7.3 Classification of Fairness Standards

CATEGORY	GROUPS	INDIVIDUALS
Interpretation/use	7.1, 7.2, 7.8, 7.9 (system) 7.10 (evidence), 7.11	7.5
Statistical bias	7.1*, 7.3, 7.6	7.7 (language)
Sensitivity	7.4	7.12

degree to which interpretations of scores on a given test are consistent with the construct. How does one frame the question about what a test measures? This is more complicated than it might seem because typically two different kinds of interpretations are involved. First, extrapolations might be made beyond the particular test items occurring on (say) a mathematics test. Test users seek to know something about a person's competence relative to the domain of all relevant test questions, and the definition of this domain partly identifies the test construct. Second, inferences may be made beyond the item domain to a broader set of desirable mathematical *behaviors*; for example, students who score high on the test should be able to solve job-related problems using statistics and probability, to solve novel problems, to use math to model real-world phenomena, and so forth. The preponderance of the latter behaviors is external to any test, and consequently there is no failsafe method for identifying the most pertinent of these for anchoring inferences. Interpretive preferences are intuitively linked to expectations about what would be the case with students who receive various levels of scores on such a test.

The terminology for denoting constructs in educational testing has fluctuated over the last century. With the advent of intelligence testing, the belief that IQ was a stable characteristic of an individual that was independent of the environment led to the use of synonyms such as ability, aptitude, cognitive ability, mental ability, learning ability, and the ubiquitous "trait." Typically, schools and organization concerned with learning distinguish intelligence and aptitude from achievement. Synonyms for achievement as defined by a construct can include skill, proficiency, attainment, competence or competency, and knowledge. The more nuanced term "developed ability" has been used to acknowledge that abilities are developed by instructional and social practices in the same way that language proficiency is developed. In this sense, the meaning of "ability" is similar to that of "problem-solving proficiency." Not all synonyms work equally well in a given context. Conversely, the distinction between "traits" and "proficiencies" is important from both semantic and social perspectives because the term "proficiency" is often more appropriate for describing learning outcomes. If achievement is being measured, then a score is less aptly described as an ability or aptitude. Since a test can only demonstrate a person's concurrent state, using labels that imply immutability, or even long-term stability, can be dangerous.

3.3.2. Consequences of Testing

Supportable interpretations of tests depend on, but are not limited to, construct definition. One can think of the construct as a nexus of propositions: a student with a higher score on the test will on average do better on homework assignments; lower scores imply disruptions in the learning process that can be identified; higher scores indicate higher levels of initial preparation; reading problems portend problems in mathematics; and so on. These propositions can be used to generate testable hypotheses for the purpose of supporting or disconfirming the validity of test score interpretation.

Yet it is important to distinguish construct-relevant from construct irrelevant variance. If, for example, a great deal of baseball questions on a test of mathematical operations would discourage baseball nonenthusiasts from their best effort, then the consequences of such item content would distort the interpretation of test scores because they would partly reflect enthusiasm (construct-irrelevant variance) for baseball as well as mathematical operations. Similarly, scores from a very long test would partly reflect endurance or test-taking speed. Such extraneous influences may adversely affect the interpretations of test performance. Messick (1994) summarized,

> it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct under-representation or construct irrelevant variance. (p. 8)

Still other types of hypotheses fall farther outside the original bundle of construct-related propositions. For example, if a student receives inadequate instruction, and as a result, does not pass a test, an accurate conclusion would be that the student does not know the material. Without additional evidence, other interpretations are unsupportable: the student failed to learn; the student was unmotivated; or the student lacked the requisite aptitude. All of the latter are characterized by at least one causal attribution external to and independent of the test construct.

Suppose a testing program is claimed to be successful in promoting achievement. The problem here is that attributing causality to the test itself ("The test increases student achievement") enlarges the construct. Such claims may require serious consideration because there may not be a single perspective for defining a test construct. The previous paragraph implicitly assumed the existence of an indisputable or privileged position for delimiting the appropriate bundle of construct-relevant propositions. Yet authority for determining what is within and outside the realm of the construct is often a shared responsibility, one that requires debate if not consensus. A one-dimensional perspective of "construct-irrelevant" variance falsely suggests a sharp dichotomy between the responsibilities of those who develop or understand assessment theory, and those who use or make practical interpretations of test scores.

4. STRUCTURAL ANALYSIS OF BIAS

Statistical bias and test fairness are usually assessed by comparing item or test performance in different identifiable groups. The use of "groups" is a statistical device, used because potential bias is uncovered by aggregating evidence across test takers within such groups. The mathematical models below acknowledge that there may be individual differences within each group, but also that there may be a component of item or test variance due to group membership *per se*.

In this section, heuristic models are presented for describing statistical bias. For thrift, it is the convention herein to refer to statistical bias simply as bias, which can

be assessed logically in two distinct ways. First, there may be a question *internal* to the test "Is this item measuring the same thing for two groups relative to the other items?" Secondly, a question might involve a relationship *external* to the test "Is this item (or test) measuring the same thing for two groups relative to an independent criterion?" This independent or external criterion is often a test of criterion performance in the context of use, or another test thought to be fair—at least fairer than the test being studied. These topics are given a theoretical treatment in this section, and a more general treatment in sections 5 (external analysis) and 6 (internal analysis).

Measurement models are mathematical equations that describe components of test scores. The most common measurement model can be expressed as

$$u_{pi} = t_p + \varepsilon_{pi}. \tag{1}$$

This equation expresses the idea that an observed measurement u for person p on item i is composed of a *true score* t_{pi} (see the reliability chapter), and an error component ε_{pi} that describes an effect due to a particular combination of person and item. Equation (1) is a theoretical notion of measurement because in actuality only the variable u_{pi} is known, where u in this case represents a scored item response.

4.1. Differential Item Functioning (DIF)

While equation (1) provides a conceptual basis for measurement, an elaboration is necessary for modeling differential group measurement:

$$z_{gpi} = \alpha_{gi}\theta_{gp} - \delta_{gi} + \varepsilon_{gpi}. \tag{2}$$

Here, z_{gpi} signifies the propensity of a person p from group g to answer item i correctly, and θ_{gp} is a latent proficiency as defined in item response theory. This expression includes an intercept δ_{gi} representing item difficulty (higher values indicate higher difficulty), a discrimination coefficient α_{gi}, and measurement error ε_{gpi}, where the subscript g indicates that a coefficient can differ for two groups (R and F). Bias in item difficulty is then expressed as $\delta_{ri} \neq \delta_{pi}$, while bias in item discrimination is obtained as $\alpha_{ri} \neq \alpha_{fi}$.

Assumptions of linearity and additivity above are made to facilitate conceptual understanding. In practice, the nonlinear expression

$$P(u_{gpi}=1) = \frac{\exp\left(\alpha_{gi}\theta_{gp} - \delta_{gi}\right)}{1 + \exp\left(\alpha_{gi}\theta_{gp} - \delta_{gi}\right)} \tag{3}$$

based on item response theory (IRT) is more common. Here, $P(u_{gpi} = 1)$ is the probability of a correct response on a dichotomous item, and the expected value of z_{gpi} in equation (2) is defined by the logit link function

$$E\left[\, z_{gpi} \,\right] = \ln\left(\frac{P(u_{gpi}=1)}{1 - P(u_{gpi}=1)} \right). \tag{4}$$

When $\delta_{ri} \neq \delta_{fi}$ or $\alpha_{ri} \neq \alpha_{fi}$ is the case, the item is said to function differently for two groups, or, alternatively, to show DIF. Because DIF is based solely on the item scores and group indicators, this is an internal (to the test) analysis of bias.

4.2. Differential Prediction

In some applications, the goal is to use a test score to predict a criterion of interest, such as using an evaluation of student teaching experience to predict supervisory ratings of first-year teaching performance. The measurement model in equation (1) omits reference to such an *external* variable (i.e., actual teaching performance), and this situation can be addressed by designating a target variable or *criterion* (say y). This choice may depend on a number of reasons, and more than one plausible criterion might exist. Given an explicit choice, predictive bias can be defined, parallel to equation (2), in terms of

$$y_{gp} = a_g x_{gp} + d_g + e_{gp}\,, \tag{5}$$

where x_{gp} is a total score on the predictors test (commonly the sum of scored item responses).

The term a_g in this equation denotes the slope of the regression model, and d_g is the model intercept. Parallel to bias in item parameters, two types of bias also exist in this situation. If the term a_g (or d_g) is not equal for two groups, the bias is present in the model in terms of slopes (or intercepts). Because this approach requires a criterion measurement y independent of the test in question, it is an external analysis of bias.

4.3. Integrated Structural Diagram

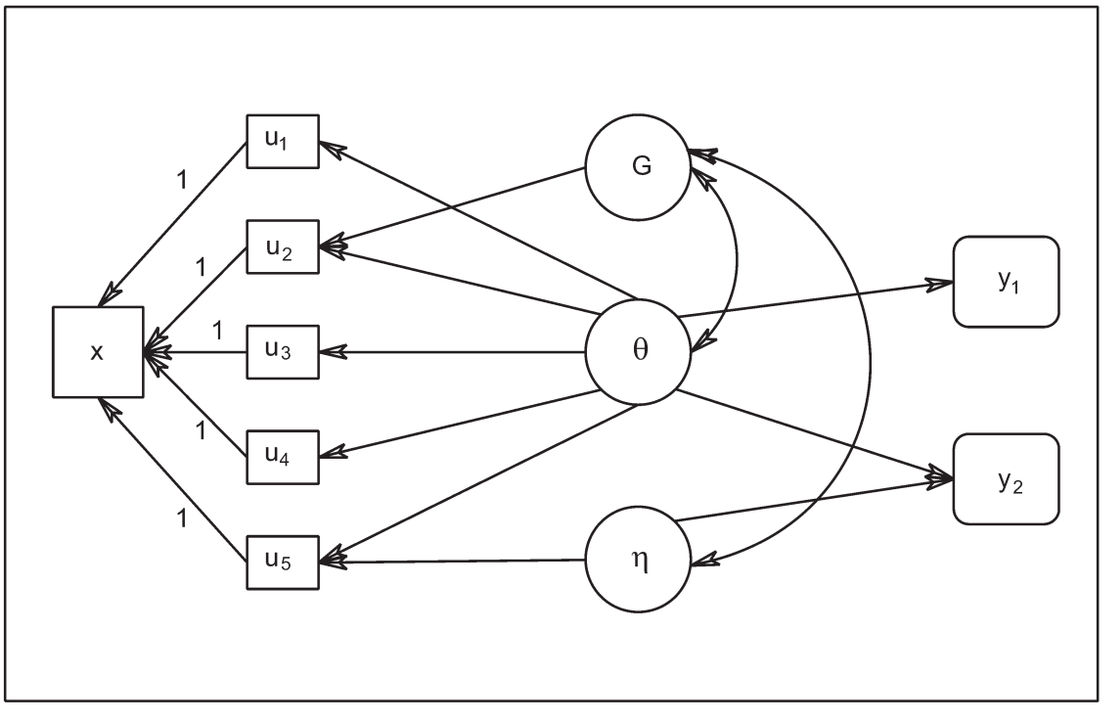
In Figure 7.1, a diagram for statistical bias is given for a scenario consisting of a hypothetical five-item test in which three variables can affect item performance. Two of these are proficiency scores, θ and η, where θ is the intended or target construct and η is a secondary factor. The third, G, is a dichotomous group indicator (R or F), which is correlated with θ and η (depicted by bidirectional arrows). The arrow directly from G to item 2 (u_2) represents DIF in an item that is affected by group membership beyond the two proficiencies that contribute to item performance. Item u_5 also may exhibit DIF because is it affected by G indirectly through the secondary factor η ($G \leftrightarrow \eta \rightarrow u_5$). Simply put, DIF can occur when factors other than target proficiency affect item performance, either directly or indirectly.

Figure 7.1 also contains a scenario containing a criterion y and a predictor variable x, where x is obtained as the simple sum of the scored item responses—indicated by the fixed "1s" on the paths from each u_i to x. Assuming the latent proficiencies, θ and η, would be the ideal predictors of y, but that only x is available, the central question concerns the behavior of the common regression prediction

$$\hat{y} = b_0 + b_1 G + b_2 x\,. \tag{6}$$

Because G does not contribute to the prediction of y_1 or y_2, unbiased regression coefficients are obtained with θ and η. It can be shown in this case mathematically that using the observed score x as a proxy for latent proficiency results in

FIGURE 7.1 A Structural Model Illustration with Five Items for Depicting Item and Test Bias



Error terms are not shown. Directional arrows represent causal effects; bidirectional arrows represent covariances.

$b_1 \neq 0$, due to DIF in u_2 and u_5. Thus, the presence of DIF in this heuristic diagram results in differential prediction, but there are two important qualifications to this statement. First, the structural model could be modified to add a direct path from G to y_1 or y_2, and this would create a scenario in which both measurement bias and predictive bias simultaneously exist. This signifies that predictive bias is not necessarily caused by DIF. Second, the secondary factor η may be relevant to the predicted outcome as in y_2. If this is the case, predictive bias due to DIF is diminished because η validly contributes to both u_5 and the criterion measurement.

In this conceptual approach, statistical bias occurs as a function of G. Bias in the sense of either differential measurement or prediction is not a property of the test *per se*, but rather a property of test use with the particular examinee populations constituting G.

5. EXTERNAL EVIDENCE OF BIAS

External bias detection requires a construct by which a person or student can be deemed successful in a given activity, and is often motivated by the need to make distinctions among examinees *before* they engage in the activity in question. The idea is to *select* examinees using their scores on the *predictor* variable. In Figure 7.2, an example of an unbiased test/criterion regression is illustrated. Following the conventions above, the two groups are labeled R and F. Despite the difference in group distributions on the predictor, the test has equal predictive validity for two groups—they share the same regression line. For any given test score, individuals have the same expected criterion performance, regardless of group membership. Figure 7.3 provides an example of a criterion-predictor relationship that differs for two groups are equal on the criterion but vastly different on the predictor.

FIGURE 7.2 Depiction of Unbiased Prediction for Two Groups R and F

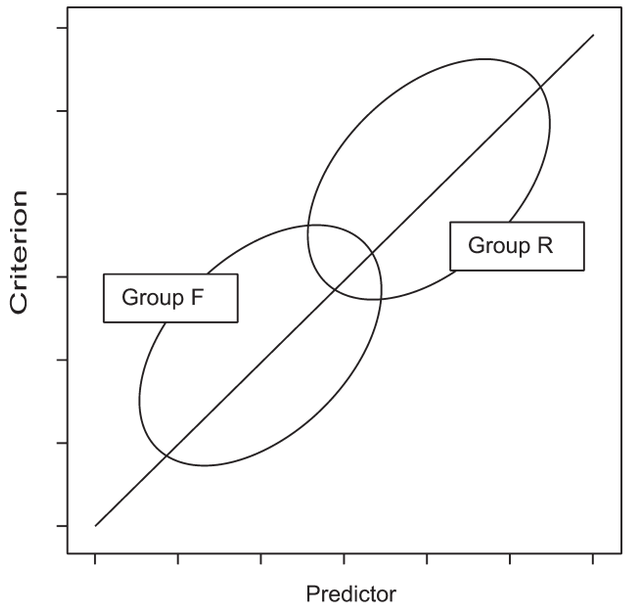
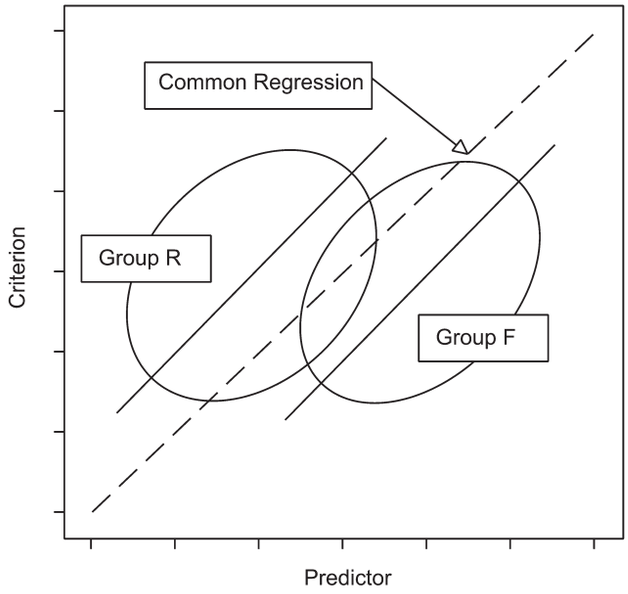


FIGURE 7.3 Depiction of Biased Prediction for Two Groups R and F



Solid lines are individual group regressions. The dotted line represents the common regression line.

In the latter case, the test differentially predicts the chosen criterion *performance* because there are two different regression lines. Although the regression slopes are the same for both groups, the regression line for the group R has a higher intercept. If a common regression line (dotted line) were used to select candidates for college or a job, the test would lead to overprediction of performance for candidates from group F leading to biased selection. Though the phrase *test bias* is often used to describe this situation, the phrase *differential prediction* more accurately conveys the meaning, and the term *fairness* is bound to a particular use of a test in selecting candidates. It is the selection procedure that is described as fair or unfair, not the test itself.

5.1. Selection Definitions

A number of technical definitions and conventions are commonly used in the analysis of selection procedures, and these are provided in Table 7.4. These terms are used extensively in the following section, and are presented here for easy reference.

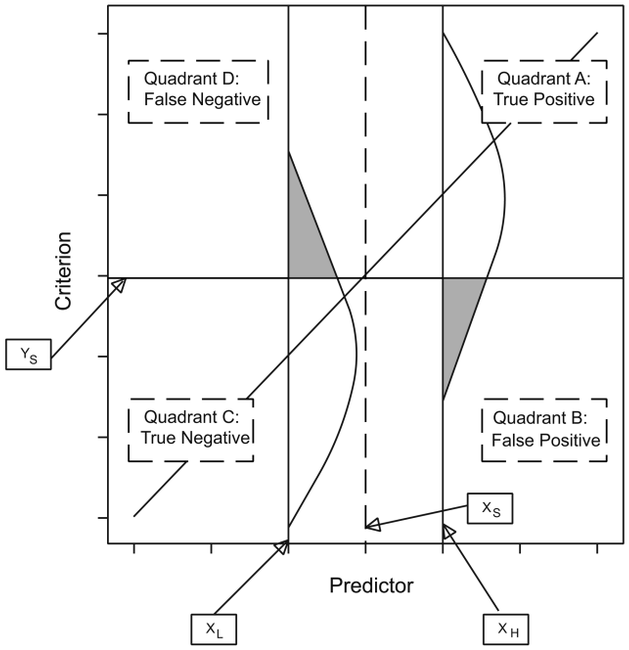
A number of potential selection models exist, each having different definitions for and implications about the fairness of the selection process. The conceptual issue these models address differently is that even with unbiased prediction, two candidates from different groups *who could perform equally well on the criterion* may not have the same probability of being selected (Thorndike, 1971). In Figure 7.4, an illustration of selection is given for an unbiased test. One set of candidates has a higher value of the predictor (X_H), and one has a lower value (X_L). If there are many candidates at each score level, a normal distribution can be used to describe the range of scores (on the vertical dimension of the graph) *within* the classes obtaining scores of X_H and X_L. These conditional distributions arise because the criterion is predicted with error.

Now suppose a score X_S is chosen on the X-axis as a cut point for selection, and only examinees with scores equal to or greater than X_S are chosen. The score Y_S denotes the criterion score of an individual that fell precisely at the predictor cut score. Upon examining the distribution of Y for X_H and X_L, an irony of the selection model is evident: some candidates at X_L score higher than Y_S and some candidates at X_H score lower than Y_S. In Figure 7.4, the shaded area above Y_S for candidates at X_L represents the probability of success (false negatives), while the shaded area for candidates at X_H represents the probability of failure (false positives). Given a less-than-perfect correlation between predictor and criterion, any reasonable cut point would result in some selected candidates being less successful than some rejected candidates would have been. As the correlation between predictor and criterion decreases, the problem is that false negative errors occur at a higher rate for the lower-scoring group. Though the test treats *individuals*

TABLE 7.4 Selection Terminology. The Numbers Falling Into Each Quadrant Are Given by A, B, C, and D as Shown in Figure 7.7, Where N = A + B + C + D

Term	Definition	Statistic
Predictor cut score	Select if $X > X_S$	X_S
Criterion cut score	Success if $Y > Y_S$	Y_S
	Number of cases selected	A + B
	Number of cases rejected	C + D
	Number of successful cases	A + C
	Number of unsuccessful cases	B + D
Base rate	proportion of cases deemed “successful” on criterion	(A + D) / N
Selection ratio	proportion of cases selected	(A + B) / N
Success ratio	proportion of selected cases that “succeed” on criterion	A / (A + B)
False positive error	a case predicted to succeed on criterion, but does not	(any case in quadrant B)
False negative error	a case predicted not to succeed on criterion, but does	(any case in quadrant D)

FIGURE 7.4 Unbiased Regression Showing Probabilities of Success at X_L and Failure at X_H



equivalently, regardless of group membership, lower-scoring *groups* bear the burden of the test's fallibility.

Being successful (as in college) is not a lucky happenstance, but depends on effort and achievement. Thus, it is a matter of concern that some rejected candidates would have been successful if they had the opportunity, and this concern leads to two central issues in the analysis of test fairness. First, the choice of criterion construct is of paramount importance. In situations with low or modest correlations of predictor to criterion, the construct of the selection test and errors of measurement drive the selection process more than the construct defining the criterion. The predictor may fail to take into account the genuine factors that lead to success such as motivation, study habits, academic support, and so forth. Second, an elite group would be selected in the case of low predictor correlation, but such elite status would be weakly related to success as defined by the criterion.

Wightman (2003) illustrated the first point with respect to the LSAT:

> The most compelling aspect of the bar admission data is that between 88 and 72 percent of minority law school students who would have been denied opportunity to enter law school under a numbers-only admission model were able to successfully pass the bar and enter the profession. Similar studies in other educational settings should be undertaken to help put the impact of selection based disproportionately on test score results into perspective. (p. 20)

Similarly, Chung-Yan and Cronshaw (2002) found that cognitive ability tests used in employment decisions typically showed a one standard deviation (SD) difference for whites and African Americans. The difference on actual job performance measures shrank to 1/3 SD; and when objective tests were used, rather than subjective rating criteria, the difference shrank to 1/10 SD. Therefore, even if the test measures a competence equally well for two groups in a particular selection process, it cannot be said *unambiguously* that the test use system is fair: two individuals from different groups, with equally likely probabilities of success, may have unequal probabilities of selection. The selection process is fair relative to the predictor, but not (retrospectively) in terms of performance. On the other hand, being fair to a group creates a mechanism in which some individuals with lower scores on the predictor will be chosen over others with higher scores.

5.2. Selection Models

The most widely recognized selection models are briefly reviewed below. These models can be conceptualized in terms of the selection model shown in Figure 7.4 and definitions given in Table 7.4. Additional references for this material can be found in the special spring 1976 issue of the *Journal of Educational Measurement*, Jensen (1980), and Schmidt and Hunter (1998).

5.2.1. Regression Model

In this model, cases having highest predicted criterion scores are selected, regardless of group membership. If groups have different regressions, then these separate regressions are used to make the predictions. This model maximizes the mean criterion score of selected cases relative to other selection models. A probabilistic model called the *equal risk model* (Einhorn & Bass, 1971) bears a strong conceptual similarity to the regression model.

5.2.2. Equal Risk Model (Petersen & Novick, 1976)

In this procedure, a threshold, say Z, for the probability of success (or, conversely, risk) in terms of the criterion score is chosen. The predictor cut score for any of K groups is then chosen so that

$$\text{Prop}(Y \geq y_s \mid X = x_{sg}, \textit{Group} = g) = Z \qquad (7)$$

for $g = 1, 2, 3, \ldots, K$. This model provides the same cut scores regardless of whether Z is defined in terms of risk or in terms of probability of success. For some selection models, it is problematic that this apparently trivial choice in defining Z results in different cut scores.

5.2.3. Constant Ratio Model (Thorndike, 1971)

Cut scores are set so that the number of selected members of a group (A + B) is proportional to the number who succeed (A + D). For this to occur, the fraction (A + B)/(A + D) must be the same for all groups.

5.2.4. Conditional Probability Model (Cole, 1973; Darlington, 1971)

This has also been described as the equal opportunity model (Wigdor & Sackett, 1993). Cut scores are set so that for candidates who would be successful (A + D), the proportion selected is the same for all groups. For this to occur, the ratio A/(A + D) must be the same for all groups.

5.2.5. Equal Probability Model (Linn, 1973)

Cut scores are set so that for candidates who would be selected (A + B), the proportion who would be successful is the same for all groups. For this to occur, the ratio A/(A + B) must be the same for all groups.

5.2.6. Decision-Theoretic Model (Petersen & Novick, 1976)

In this model, a utility (a quantitative defined benefit to individuals or society) must be assigned to each outcome: true positives, true negative, false positives, and false negatives. Higher utilities are operationalized as a higher weight assigned to any outcome "that has the effect of equalizing opportunity or reducing disadvantage" (Novick & Petersen, 1976, p. 83). A cut score for two or more groups is determined as the value providing the greatest benefit across individuals.

5.3. Summary of Models

The National Research Council Committee on the General Aptitude Test Battery (Hartigan & Wigdor, 1989) categorized selection models into two sets. Methods (5.2.3–5.3.5) that take into account the number of successful candidates are described as "performance fair" in contrast to the rule of the regression and equal risk models (5.2.1–5.2.2), which can be described as "test fair." The former take into account group membership, whereas the latter only recognize individual performance on the predictor test (unqualified individualism). The decision-theoretic model (5.2.6) recognizes that selection takes place in a context of values, and an attempt is made to make these values explicit in the decision rule. This is clearly a desirable approach, yet it is difficult to implement because there may be no consensus for determining the required utilities. Models 5.2.3–5.2.5 attempt to take performance into account in a simpler manner, but these models are internally inconsistent. Depending on whether one seeks to minimize risk or maximize utility, different cut scores are obtained with the identical decision logic (Petersen & Novick, 1976).

Group fairness is not the same as individual fairness, and debate arises regarding quota systems and preferential treatment whenever methods are suggested to correct for the disproportionate effect of fallible tests on protected groups. Jensen (1980) argued that any model other than unqualified individualism is a *quota* method, and the common criticisms of quotas in selection processes are three-fold. First, a quota method runs counter to the American ideal of individual fairness. Under a quota system the rules of play are modified to favor one or more groups of individuals by arbitrarily increasing their likelihood of selection. Second, such methods are criticized as racist because they perpetuate the stereotype that protected group status is correlated with potential for success. The third criticism is that only the individual model "maximizes" utility (Jensen, 1980). In this case, unqualified individualism is the preferred selection procedure; accordingly, scarce resources are allocated to the most talented applicants, who return the highest level of benefit to society.

As a whole, these criticisms constitute less than a complete argument. Because protected groups often have less "access to privilege, power, and position" (Smedley, 2002, p. 174), tests are often used in social contexts of decidedly unbalanced playing fields. Race in this context is a proxy for access, that is, the tools for effectively competing. With respect to the selection procedure itself, fairness is not unambiguously established with a test having low to moderate correlation with a criterion measure of performance. As shown by Wightman (2000) in her analysis of the LSAT, protected status can have a large impact on selection, but has a much smaller degree of association with realized success. The criticism of maximizing utility hinges on the definition of utility. It assumes a single, or at least preferable, criterion that provides a greater benefit to society than all others and that selected applicants do indeed use their talents to benefits society, either directly or indirectly.

For these reasons, a test that can be demonstrated as statistically unbiased still may not meet the criterion of educational or economic necessity if it disparately impacts applicants. The attempt to clarify this issue can be seen in the Supreme Court decision *Grutter v. Bollinger et al.* (2003), where the majority held that the University of Michigan Law School can employ diversity as one component in formulating an admission policy:

> Enrolling a "critical mass" of minority students simply to assure some specified percentage of a particular group merely because of its race or ethnic origin would be patently unconstitutional. . . . But the Law School defines its critical mass concept by reference to the substantial, important, and laudable educational benefits that diversity is designed to produce, including cross-racial understanding and the breaking down of racial stereotypes. The Law School's claim is further bolstered by numerous expert studies and reports showing that such diversity promotes learning outcomes and better prepares students for an increasingly diverse workforce, for society, and for the legal profession. (pp. 3–4)

Empirical investigations of equal prediction can answer whether a test is performance or individual "fair" for two groups, but cannot address larger questions about the social value of selection procedures (Linn, 1989). The dissenting opinion of Justice Clarence Thomas in *Grutter v. Bollinger* illustrates the difficulty in arriving at a consensus regarding social utility:

> the Law School seeks to improve marginally the education it offers without sacrificing too much of its exclusivity and elite status. The proffered interest that the majority vindicates today, then, is not simply "diversity." Instead the

> Court upholds the use of racial discrimination as a tool to advance the Law School's interest in offering a marginally superior education while maintaining an elite institution. Unless each constituent part of this state interest is of pressing public necessity, the Law School's use of race is unconstitutional. I find each of them to fall far short of this standard. (p. 8)

Whereas the majority recognized multiple social values in admissions policies, Thomas argued that the architecture of the selection process is flawed, and consequently students are denied equal protection under the 14th Amendment. The use of developed ability as a predictor of success in combination with a stringent cutoff creates an elite law school, but one without a compelling state interest, Thomas maintained, because few Michigan students attend the law school and of these, most eventually leave the state.

5.4. Empirical Results

Empirical studies have demonstrated a tendency toward overprediction in favor of the focal group performance when a common regression approach is used. Two areas of research are summarized below to illustrate this general conclusion.

In a review of 49 separate studies completed since 1974 on college admission testing, Young (2001) found that for most African Americans and Hispanics, first-year grade point average (FGPA) was slightly *overpredicted* by college admissions test, or about .11 units on a four-point grade scale, meaning that students in these groups perform slightly worse than the test predicts. For women, FGPA was *underpredicted* by about .05 to .06 units. Thus, there is convincing empirical support that the degree of differential prediction is not large for various gender, racial, and ethnic groups on college admissions tests. Young found the average validity coefficient (multiple correlation) to be about .5. Other studies have reached similar conclusions (Linn, 1982; Ramist, Lewis, & McCamley-Jenkins, 1994; Wightman, 2003).

In a review of validity studies concerning the General Aptitude Test Battery (GATB), the GATB Committee identified 70 studies using this test to predict criterion outcomes. Hartigan and Wigdor (1989) reported,

> In 26 of the 70 studies the intercepts were significantly different at the .05 level . . . in only 1 of the 26 studies in which the intercepts were significantly different was the intercept-greater for black than for nonminority employees. (p. 181)

In this case as in the case of the SAT, the common-groups regression equation is more likely to overpredict than underpredict the performance of black applicants. However, for 72 validity studies "that had at least 50 black and 50 nonminority employees" (p. 188), the average correlation of the GATB with the criterion for the former group was $\bar{r} = .12$, and for the latter group $\bar{r} = .19$. For one quarter of the studies, Hartigan and Wigdor (1989) reported a correlation of $\bar{r} = .03$ or less.

6. INTERNAL EVIDENCE OF BIAS

Procedures using internal evidence to detect bias received considerable attention beginning in the mid-1970s. Methodologically, the intent of DIF analyses was to distinguish *bona fide* group differences from bias in the measurement process. Group differences in test performance cannot be interpreted automatically as evidence of either bias or unfairness because these differences might validly reflect construct-relevant knowledge and opportunity. Therefore, the concept of *relative* difficulty was devised (probably by William Angoff at Educational Testing Service). Absent an external criterion, a variety of internal bias procedures were developed using the other items on the test: an item of interest had a group performance difference relatively larger than the group differences for other items. The modern methods of DIF represent a refinement of this notion.

A major limitation of item bias statistics or indices is that measures of relative difficulty do not provide proof of unfairness. Only if an item is relatively more difficult for one group (statistically biased) *and* the source of this difficulty is irrelevant to the test construct is an item said to be unfair. Holland and Thayer (1988) introduced the term *differential item functioning* to convey this concept more clearly:

> The study of items that function differently for two groups has a long history. Originally called "item bias" research, modern approaches focus on the fact that different groups of examinees may react differently to the same test question. These differences are worth exploring since they may shed light both on the test question and on the experiences and backgrounds of the different groups of examinees. We prefer the more neutral terms, differential item performance or *d*ifferential *i*tem *f*unctioning, (i.e., *dif*), to item bias since in many examples of items that exhibit *dif* the term "bias" does not accurately define the situation. (p. 129)

In other words, DIF is synonymous with statistical bias, whereas unfairness can only be established if these measurement differences are factors irrelevant to the test construct; there is no direct route from statistical bias to unfairness. To maintain the distinction between statistical bias and unfairness, DIF is used as one kind of screening mechanism for quality control. The process of using such an index jointly with a logical analysis of potential attributions is a procedure for *detecting item unfairness.* The term DIF is now widely used in the literature, but, unfortunately, some writers still suggest DIF is sufficient for detecting *unfair* items. Ideally, DIF statistics are used to identify all items that function differently for different groups; then, after logical analysis as to *why* the items seem to be relatively more difficult, the subset of DIF items identified as "unfair" would be eliminated from the test.

Findings from item bias analyses may also help to clarify what a test is measuring and highlight the influence of irrelevant factors. Most major achievement tests have a single dominant or "essential" factor or dimension (Reckase, Ackerman, & Carlson, 1988; Shealy & Stout, 1993b). Because DIF statistics work by signaling systematic group differences, they are highly sensitive to multidimensionality when a secondary dimension is relevant to answering

611-190-007.indd 234 7/14/2006 4:18:21 PM

an item correctly *and* when groups differ on one or more secondary dimensions. Therefore, DIF analyses provide insights much like an item-level factor analysis. If a secondary factor is identified in what was believed to be a homogeneous measure of a single proficiency, then the test developer is forced to consider explicitly whether the secondary proficiency is an admissible part of the intended construct. For example, Shepard, Camilli, and Williams (1984) found that verbal math problems were systematically more difficult for African American examinees; differences between this group and Caucasians were larger on this type of problem than on straight computational problems. In this case, findings from the DIF screening might prompt a more conscious appraisal of what proportion of test items should be word problems.

6.1. An Historical Caution

Before describing current methods of DIF analysis, it is useful to reconsider an earlier rationale for defining and operationalizing these procedures. Eells, Davis, Havighurst, Herrick, and Tyler (1951) were not the first researchers to address the question of socioeconomic differences on intelligence test items; however, they drew together much of the literature—nine studies from 1911 to 1947 were reviewed—and performed a primary analysis of more than 650 items from eight IQ tests. Moreover, they defined *cultural bias in test items* as

> differences in the extent to which the child being tested has the opportunity to know and become familiar with the specific subject matter or specific process required by the test item. If a test item requires, for example, familiarity with symphony instruments, those children who have opportunity to attend symphony concerts will presumably be able to answer the question more readily than children who have never seen a symphony orchestra. (p. 58)

Eells et al. were interested in establishing the extent to which observed group differences in IQ scores were dependent on the specific content of the test items rather than an important underlying thinking ability in pupils (p. 4). A test was considered fair if it was composed of items that were equally familiar or unfamiliar to all persons.

The two major purposes of the Eells et al. (1951) study were to detect differential measurement, and then to discover "a) those kinds of test problems on which children from high socioeconomic backgrounds show the greatest superiority and b) those kinds of test problems on which children from low socioeconomic backgrounds do relatively well" (p. 6). This knowledge then would be used to eliminate *cultural bias,* favoring any particular socioeconomic group, from the test (p. 24). When items showing large group differences in performance were detected and analyzed for commonalities, Eells et al. (1951, p. 68) concluded that "variations in opportunity for familiarity with specific cultural words, objects, or processes required for answering the test items seem to the writer to appear…, as the most adequate general explanation for most of the findings."

By eliminating items that relied on opportunity to learn, Eells et al. (1951) believed that group differences then would reflect more accurately an important underlying ability. In 1951 things did not work out so neatly. Eells et al. (1951) concluded their study with the following caution:

> Another important finding in the analysis reported in this chapter is the rather substantial number of items showing large status difference for which no reasonable explanation can be seen. (p. 357)

Thus, about a half century ago analysts had noted that what would later be known as differential item functioning often had no satisfactory explanation. While it is true that the early methods for DIF were unsophisticated statistically, it is also true that findings with no apparent explanation are the same general complaint of analysts today. For example, in studies in which expert judges try to identify DIF-prone items based on substantive or qualitative criteria, little correlation has been found between expert ratings and empirical DIF indices (Englehard, 1989; Reynolds, 1982). Group membership itself, especially with regard to race, is an inference loaded with implicit and usually untestable assumptions, it is not surprising that the observed statistical relationships with this variable have been inconsistent, if not bewildering. As Bond (1994) stated,

> Theories about why items behave differently across groups can be described only as primitive. Part of the problem, as I see it, is that the very notion of differential item functioning by groups implies a homogenous set of life experiences on the part of focal groups that are qualitatively different from the reference group that affect verbal and mathematical reasoning. As I have indicated elsewhere (Bond, 1980, 1981) we are still far from a coherent theory of how knowledge is organized in long-term memory, how it is retrieved, and how it is used in problem solving. (p. 279)

However, some narrow regularities have been noted in DIF research. For an example of substantive findings regarding language, see Schmitt, Holland, and Dorans (1993).

6.2. Other Early DIF Studies

The first well-known studies of test item functioning began in 1910 with Alfred Binet, who was concerned that some items may disadvantage students from lower socioeconomic strata. Binet eliminated certain kinds of items that were considered too dependent on home training or language. William Stern in 1912 also observed such item differences and recommended that procedures be developed for detecting such items. Thurstone (1925, 1931) conducted related studies in which items were examined for group changes in relative scale positions.

Until the 1960s, investigations like that of Eells et al. (1951) tended to focus on items from intelligence tests. More modern investigations began with Coffman (1961), and Cardall and Coffman (1964), who wrote that

> a comparison of the responses of different groups of subjects to a set of test items generates two statistics of interest. First, one may ask whether or not there are mean difference across groups, that is whether the groups differ in average

> test score. Second, one may ask whether there is a significant interaction between the items and the groups, that is, whether or not some particular items are relatively easier for one group than for another. (p. 2)

Continuing this work, Angoff and Ford (1973) studied race by item interaction on the Preliminary Scholastic Aptitude Test (PSAT) by matching candidates on total mathematics or verbal subscores. They recommended multivariate matching strategies for DIF investigation, and demonstrated the use of "delta plots," a currently popular graphical procedure (see Camilli & Shepard, 1994, for a fuller description of this method). In the late 1970s and early 1980s, DIF research began to flourish (e.g., Berk, 1982; Ironson & Subkoviak, 1979; Lord, 1977; Scheuneman, 1979; Shepard, Camilli, & Averill, 1981).

7. METHODS OF DIF ANALYSIS

There are two major approaches to DIF analysis. One approach is the use of item response (IRT) theory models, while the other approach relies on methods of observed-score analysis. Both approaches share a core concept: DIF is defined as item performance differences between examinees of comparable proficiency. While IRT methods provide useful results when the item models fit the data *and* a sufficient sample size exists for obtaining accurate estimates of IRT parameters, observed-score methods are frequently used with smaller sample sizes. However, the latter methods contain implicit measurement models, so it is imprecise to refer to them as *nonparametric*. The accuracy of their results depends on how well this implicit model fits the data (Camilli & Shepard, 1994). For example, the Mantel-Haenszel technique (described below) provides an unbiased measure of DIF for a multiple-choice item only when several strong assumptions are satisfied, one of which is that a Rasch model fits the item responses.

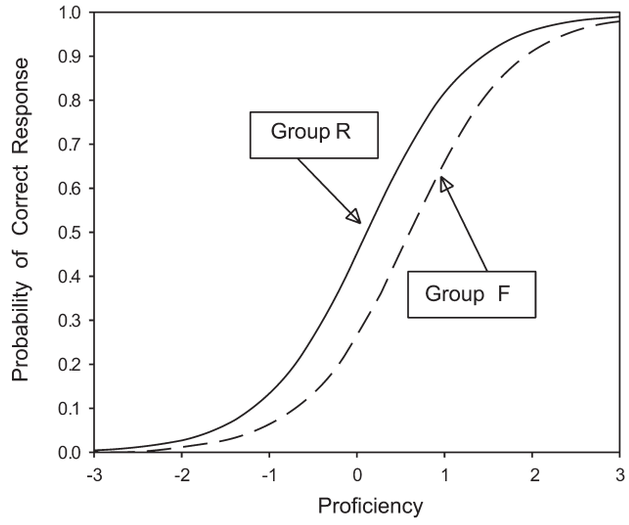
7.1. Difference in IRT Parameters

One general approach to assessing the difference in item response functions (IRFs) is a comparison of item parameters. There are various procedures, but the essential feature is the comparison of one of more item parameters across two groups (Lord, 1980). For example, under the 1-paramater logistic model, a test of the of *b*-parameter difference, H_0:$b = 0$, uses an normal approximation test statistic

$$Z = \frac{\Delta\hat{b}}{s(\hat{b})}, \tag{8}$$

where $\Delta\hat{b} = \hat{b}_F - \hat{b}_R$ and $s(\hat{b})$ is the standard error of the difference. This situation is illustrated in Figure 7.5. Since the item response functions are parallel, this is an instance of *uniform* DIF. For polytomous items, it is convenient to express the item difficulty in terms of an average difficulty parameter and category deviations from this average. Then DIF can be expressed as a shift of the average difficulty parameter for two groups (Muraki, 1999) using equation (8).

FIGURE 7.5 Uniform DIF Expressed as the Difference in *b* Parameters for the Item Response Functions (IRFs) for a Reference and Focal Group



7.2. Difference in Item Response Functions

When a 2- or 3-PL model is used, a multivariate test of item parameter differences may be more appropriate (Camilli & Shepard, 1994; Hambleton & Swaminathan, 1984). In this case, the test is between IRFs rather than individual IRT parameter estimates. Three general approaches have been taken.

7.2.1. Multivariate Difference in Parameters

Let the vector of item parameter differences be expressed as

$$V = (\hat{a}_F - \hat{a}_R, \hat{b}_F - \hat{b}_R, \hat{c}_F - \hat{c}_R). \tag{9}$$

The statistic for testing item bias, referred to as Lord's chi-square, is given by

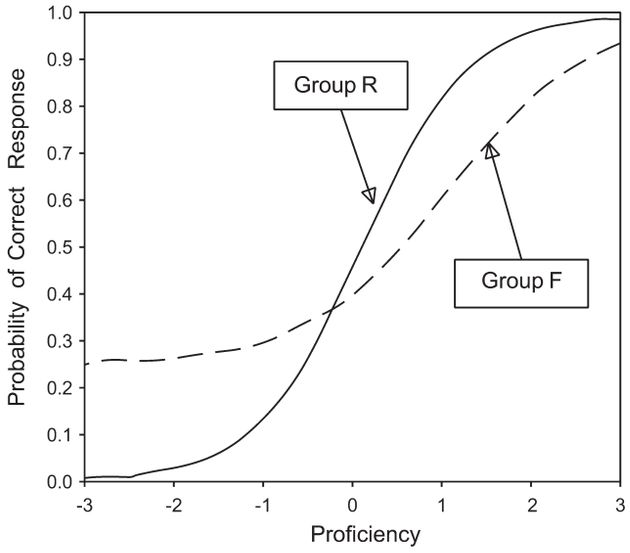
$$Q = V S^{-1} V', \tag{10}$$

where S is the sample variance-covariance matrix of the differences between the item parameters (Lord, 1980). The statistic Q is distributed as chi-square random variable with degrees of freedom equal to the number of parameters estimated. This case is illustrated in Figure 7.6. As can be seen, the amount of DIF varies by level of proficiency; this phenomenon is known as *nonuniform* DIF.

7.2.2. Area Between IRFs

In simple area measures, DIF is indicated by the area between the IRFs and may be signed or unsigned (Camilli & Shepard, 1994; Raju, 1988). In both cases, the smaller the area, the less DIF. Mathematically these area measures are defined as closed-form solutions to the integrals:

FIGURE 7.6 Nonuniform DIF Expressed as the Difference in Item Response Functions (IRF) for a Reference and Focal Group



$$\text{Signed-Area} = \int_{-\infty}^{\infty} [P_R(\theta) - P_F(\theta)]\, d\theta, \tag{11}$$

and

$$\text{Unsigned-Area} = \int_{-\infty}^{\infty} |P_R(\theta) - P_F(\theta)|\, d\theta, \tag{12}$$

where $P_G(\theta)$ is the IRF for group G. Both of these measures are effect sizes. If the Signed-Area is positive, then the Reference group is favored. In the case of nonuniform DIF, the Signed-Area would tend to cancel, thus reducing the measure of DIF. In this case, the Unsigned-Area is preferable. A large discrepancy between the two area measures indicates crossing IRFs. Raju (1990) provided asymptotic formulas for the standard errors of equations (11) and (12), which can be used for obtaining statistical tests. The major weakness of this approach is that these measures may be distorted by IRF differences in sparse regions of the θ continuum. In addition, when the c parameters differ, the integrals do not yield finite values (Camilli & Shepard, 1994).

7.2.3. Likelihood Ratio Test

In this procedure, the fit of an augmented model (A) in which IRT parameters of a studied item are allowed to vary across comparison groups is compared to the fit of a compact model (C) in which item parameters are constrained to be equal across groups (Thissen, Steinberg, & Wainer, 1988, 1993). The likelihood L of the compact model is calculated during a single calibration run. Subsequent runs calculate the likelihood of an augmented model (A) that relaxes some combination of a, b, and c parameters, for a single item for the focal group. This is equivalent to treating the studied item as a different item in each group. The null hypothesis, H_o: Model C (no DIF) is rejected in favor of the alternative, H_a: Model A, using the likelihood ratio test that the simpler model holds. The likelihood ratio test, given by

$$G^2(df) = -2 \ln \left[\frac{L(A)}{L(C)} \right], \tag{13}$$

has a large-sample chi-square distribution with degrees of freedom equal to the difference in number of parameters in the two models (Thissen et al., 1993).

Item response theory DIF methods using 2- and 3-PL models require larger sample sizes, and thus may not be appropriate when one of the comparison groups is relatively small (Clauser & Mazor, 1998). With all three parameters relaxed, the difference between IRFs is tested, and this may be preferable to testing parameters individually because very different combinations of parameters can result in similar IRFs. The likelihood ratio procedure is easily extended to polytomous items, whereas no such extensions of the procedures in 7.2.1 and 7.2.2 have been reported in the literature.

7.3. Observed-Score Methods: Mantel-Haenszel Statistics

Observed score techniques provide an alternative to IRT-based procedures when sample sizes are small or when strong assumptions can be made about underlying models. The most widely used, the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959) was introduced to psychometrics by Holland and Thayer (1988) to study group differences on dichotomously scored items. Applied to test data, the MH procedure pools information across levels (say j) of the matching variable, based on the assumption of a common odds ratio. The resulting estimate is interpreted as the relative likelihood of success on a particular item for comparable members of two different groups of examinees. This odds ratio provides an estimated effect size, and a value of 1.0 (equal odds) indicates no DIF (Dorans & Holland, 1993).

The typical setup for this analysis is a $2 \times 2 \times S$ matrix of items from a K-item test that are dichotomously scored (say 0, 1). The matching variable describing the third dimension of this matrix is typically taken as the total test or number right score ($j = 1$ to $S - 1$). For each level j, the studied item responses are tallied by item score (0, 1) and group (reference, focal) giving the matrix in Figure 7.7. An effect size measure of DIF is then obtained as the Mantel-Haenszel odds ratio:

$$\alpha_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}. \tag{14}$$

This is typically converted to the log-odds scale by the transformation $\hat{\delta}_{MH} = \ln(\alpha_{MH})$ with approximate variance (Holland & Thayer, 1988; Phillips & Holland, 1987) given as

$$SE\{\hat{\delta}_{MH}\} = \left(\frac{1}{2U^2} \sum_j T_j^{-2} \{ A_j D_j + \hat{\alpha}_{MH} B_j C_j \} x \right.$$
$$\left. \{ A_j + D_j + \hat{\alpha}_{MH} (B_j + C_j) \} \right)^{\frac{1}{2}}, \; U = \sum_j A_j D_j / T_j. \tag{15}$$

The log-odds ratio and its associated standard error can be converted to the delta scale (used at Educational Testing Service) with the transformations

$$\text{MH D-DIF} = \Delta_{MH} = -2.35 \cdot \hat{\delta}_{MH},$$

$$SE(\Delta_{MH}) = 2.35 \cdot SE(\hat{\delta}_{MH}). \quad (16)$$

The log-odds and its MH D-DIF cousin take on values from negative to positive infinity. Negative values indicate a higher likelihood of success in the reference group; and positive values indicate a higher likelihood of success in the focal group. The associated null hypothesis is $H_o: \delta_{MH} = 0$, where $\hat{\delta}_{MH}$ is an estimate of the common log-odds across levels of the matching variable. An associated test statistic can be obtained as the MH χ^2, a chi-square random variable with 1 degree of freedom. Alternatively, a z statistic can be obtained by dividing the log-odds estimate by its standard error:

$$z = \frac{\Delta_{MH}}{SE(\Delta_{MH})}. \quad (17)$$

Zieky (1993) described three categories of DIF magnitude, labeled A, B, and C. Both χ^2_{MH} (with $p = .05$) and the absolute value of MH D-DIF are used for this classification:

- A items have Δ_{MH} not significantly different from zero, or $|\Delta_{MH}| < 1$
- B items have Δ_{MH} significantly different from zero and either (a) $|\Delta_{MH}| < 1.5$, or (b) Δ_{MH} not significantly different from 1
- C items have Δ_{MH} significantly greater than 1, and $|\Delta_{MH}| \geq 1.5$

This classification has a significant impact on which item are identified for possible deletion (see section 8.2). In some testing programs type C items are always deleted. A similar classification system for polytomous items is discussed by Zwick, Thayer, and Mazzeo (1997).

A number of simulation studies have shown that Mantel-Haenszel statistics are somewhat over- or underestimated due to several factors including: a matching variable (Holland & Thayer, 1988; Swaminathan & Rogers, 1993; Zwick, 1990); multiple DIF items (Zwick et al., 1997); a nonuniform odds ratio (Clauser, Nungester, & Swaminathan, 1996); guessing (Camilli & Penfield, 1997); lack of a sufficient statistic for matching (Zwick, 1990); and sample size (Spray, 1989). A polytomous item extension of the Mantel-Haenszel procedure (Mantel, 1963) was adapted for DIF analysis by Zwick, Donoghue, and Grima (1993). Penfield (2001) showed that the generalized Mantel-Haenszel statistic (Mantel & Haenszel, 1959) provided a better assessment than multiple tests conducted on pairs of groups.

7.4. Observed-Score Methods: Logistic Regression

Swaminathan & Rogers (1990) introduced methods of DIF analysis based on logistic regression. In their approach, the matching variable is as the total or number-correct score X, but is treated as a continuous variable. Logistic regression belongs to a broad class of models known as generalized linear models or GLM (Agresti, 1996). In logistic regression, the item response is taken as a random Bernoulli variable Y_i (scored dichotomously) for individuals i with mean and variance

$$E(Y_i \mid X_i, G_i) = P_i.$$

$$Q_i = 1 - P_i, \quad (18)$$

and

$$Z_i = \ln(P_i/Q_i) = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 X_i G_i, \quad (19)$$

where P_i is the conditional proportion of individuals that endorse an item in the direction of the latent variable. A dichotomous group membership variable G is often used signifying reference and focal groups (scored 0–1).

The studied item can be evaluated using a likelihood ratio procedure. First, likelihood estimation is performed for the equation $Z_i = \beta_0 + \beta_1 X_i + \beta_2 G_i$. Next, group membership and group-by-total score interaction are added, such that $Z_i = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 X_i G_i$. The presence of nonuniform DIF is determined with likelihood ratio test (1 *df*) for improvement in model fit. Similarly, $Z_i = \beta_0 + \beta_1 X_i + \beta_2 G_i$ can be tested against $Z_i = \beta_0 + \beta_1 X_i$. The coefficient $\hat{\beta}_2$ provides an estimate of effect size that is often very similar in value to $\hat{\delta}_{MH}$. This procedure can be extended to multiple groups with the addition of dummy codes.

Logistic regression for DIF analysis is a flexible tool (Clauser & Mazor, 1998; Clauser, Nungester, & Mazor, 1996; Rogers & Swaminathan, 1993; Zumbo, 1999). This procedure typically yields DIF effect sizes highly similar to those obtained by equation (14). One advantage is that group difference on an item can be modeled with multiple variables, and this may be more efficient than multiple matching required by MH analysis (Mazor, Kanjee, & Clauser, 1995). Moreover, Miller and Spray (1993) and Camilli and Congdon (1999) showed that logistic regression can be adapted to item responses with polytomous outcomes by switching the item response and group membership variable in equation (19).

FIGURE 7.7 General Notation for the $2 \times 2 \times S$ Data Matrix

	Score on Studied Item		
Group	1	0	Total
Reference (R)	A_j	B_j	n_{Rj}
Focal (F)	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

7.5. Observed Scores: Standardized Difference

The standardization approach was introduced to DIF analysis by Dorans and Kulick (1983, 1986), who analyzed data form the $2 \times 2 \times S$ matrix by first calculating

$$\Delta p_j = p_{r,j} - p_{f,j} = \frac{A_j}{n_{rj}} - \frac{C_j}{n_{fj}}, \quad (20)$$

where $p_{r,j} = A_j/n_{r,j}$ and $p_{f,i} = C_j/n_{f,j}$ (see Figure 7.7). This is a measure of the difference in proportions correct at level j, and a weighted average can be obtained across the levels of j by

$$D = \frac{\sum_j w_j \Delta P_j}{\sum_j w_j}, \tag{21}$$

where the weighting factor can be defined in several ways. Using, $w_j = n_{f,j}$, D can be targeted to values of the matching variable most frequently obtained by focal group members. The variance of D (Dorans & Holland, 1993) is given by

$$Var(D) = P_f(1-P_f)/N_f + \sum_j n_{f,j}^2 p_{r,j}(1-p_{r,j})/(n_{r,j}N_f) \tag{22}$$

where P_f is the proportion correct observed in the focal group, and N_f is the number of examinees in the focal group. The measure in (21) is a familiar and useful measure of effect size since it can be interpreted as the average difference (reference minus focal) in proportion correct for examinees of comparable ability.

7.6. Observed Scores: SIBTEST

The methods described thus far analyze one item at a time. The Simultaneous Item Bias procedure or SIBTEST (Shealy & Stout, 1993a, 1993b), a DIF detection method motivated by multidimensional IRT, can be used to detect DIF either in single items or in bundles of items. Bundles or subsets can be constructed to some organizing principle such as test content or item format; and pooling information across items may result in more sensitive tests of group differences. Similar to other DIF techniques, a matching or conditioning variable must be constructed or chosen to create comparable subsets of examinees.

SIBTEST incorporates secondary dimensions, η (a scalar or vector), into the mathematical model for the probability of answering an item or set of items correctly, and conditions on a subtest of items presumed to measure only the target proficiency θ. The basic index for SIBTEST, β_s, represents the expected value of group difference in subtest score across the focal group ability distribution controlling for the target proficiency factors and is given by

$$B_s = \sum_j w_j \left(\overline{Y}^*_{r,j} - \overline{Y}^*_{f,j} \right) \tag{23}$$

where B_s can be interpreted as the average difference in proportion correct for dichotomously scored items, and $E[B_s] = \beta_s$. In equation (23), w_s, is the proportion of pooled reference and focal groups on the matching test, $\overline{Y}_{g,j}$ is the average item score for group g at the jth level of the matching variable, and

$$\overline{Y}^*_{g,j} = \overline{Y}_{g,j} - M_{g,j}\left[V_g(j) - V(j) \right],$$

$$M_{gj} = \frac{\overline{Y}_{g,j+1} - \overline{Y}_{g,j+1}}{V_g(j+1) - V_g(j-1)}, \tag{24}$$

where

$$V(j) = \frac{1}{2}\left[V_r(j) + V_f(j) \right]. \tag{25}$$

Finally, the estimated matching true score is given by

$$V_g(j) = \overline{X}_g + \hat{\rho}\left(X_j - \overline{X}_g \right), \tag{26}$$

where $X_i = j$ ($j = 0, 1, 2, 3 \ldots J$) is the value of the matching variable (for creating comparability), $\overline{X}_g$ is the mean of group g on the matching test, and ρ is the reliability coefficient. An important feature of the SIBTEST procedure is the Kelley-type regression adjustment (see Braun, 2006) for measurement error applied to the matching variable given in (26). This controls for score distributions that are affected by measurement error, which also tends to inflate group mean differences based on observed scores. Based on this correction, covariance adjusted estimates are obtained for the average item scores on the studied item for each group according to equation (24). In SIBTEST, the reliability is estimated by

$$\hat{\rho} = \frac{n}{n-1}\left\{ 1 - \frac{\sum_{i=1}^{n} p_i^*(1-p_i^*)}{Var(X)} \right\} \tag{27}$$

where p_i^* is the item p-value adjusted for a chance response

$$p_i^* = (p_i - c)/(1-c). \tag{28}$$

This differs from Cronbach's Alpha only by the guessing correction applied to item p values in the numerator (Zwick et al., 1997). The standard error of the estimate given in (23) is

$$sB_s = \left\{ \sum_{j=0}^{J} w_j^2 \left(\frac{s^2(Y \mid r,j)}{n_{r,j}} + \frac{s^2(Y \mid f,j)}{n_{f,j}} \right) \right\}^{\frac{1}{2}}, \tag{29}$$

where $s^2(Y|g,j)$ is the sample variance of the studies subtest scores for examinees of group g with matching test score j. An asymptotically normal test, for the no-DIF hypothesis, $H_o: \beta = 0$ is then provided by

$$z = \frac{B_s}{s(B_s)}. \tag{30}$$

In general, the type 1 and 2 error levels of this procedure have been found to be at least as accurate as those of other procedures (Bolt & Stout, 1996; Zwick et al., 1997).

SIBTEST has been primarily used to study groups of items simultaneously so that $\overline{Y}_{g,j}$ could be the average of the sum of several items. Studying several items may provide a more powerful test of DIF (i.e., DIF amplification) if each item is sensitive to the same secondary dimensions. Single items can also be examined, and the procedure has been extended to accommodate polytomously scored items, and to allow for item bundles that may exhibit either uniform or nonuniform DIF.

8. CURRENT TOPICS IN DIF ANALYSIS

Recent developments in DIF methods have yielded substantial improvements in statistical accuracy. The preponderance of studies has been methodological, rather than addressing issues of fairness directly. Examples of recent studies that apply DIF techniques to substantive problems are exemplified by Takala and Kaftandieva (2000), who studied gender differences on a Level 2 foreign language vocabulary test, and Le (1999), who studied gender DIF on

a history achievement examination. O'Neill and McPeek (1993) summarized much of the substantive research on item characteristics associated with DIF. In contrast to substantive findings, this section focuses on methodological developments.

A brief summary of the recent methodological research is useful to consider for numerous reasons, including the fact that DIF has become a major tool in state assessments. In particular, topics are considered regarding testing versus estimation, type 1 errors, item discrimination, multidimensionality, DIF as parameters versus secondary proficiencies, and item difficulty variation (see Camilli & Monfils, 2003; Camilli & Penfield, 1997; Zwick & Lewis, 1999). These topics only scratch the surface of innovation. Other research areas that are not reviewed below include DIF techniques for open-ended items (Hidalgo-Montesinos & Lopez-Pina, 2002; Zwick et al., 1997), computer-assisted testing (Walker, Beretvas, & Ackerman, 2001; Zwick, Thayer, & Wingersky, 1994, 1995), scale purification (Clauser, Mazor, & Hambleton, 1993; Zenisky, Hambleton, & Robin, 2003), latent class analysis (Cohen & Bolt, 2002; Westers & Kelderman, 1991), and cognitive studies (Gierl, Bisanz, Bisanz, & Boughton, 2003). Still others have shown that DIF has very little effect on total score distributions (Burton & Burton, 1993; Hu & Dorans, 1989; Roznowski & Reith, 1999). Several literature reviews cover these issues and estimation methods in more depth (e.g., Camilli & Shepard, 1994; Clauser & Mazor, 1998; Dorans & Holland, 1993; Millsap & Everson, 1993; Penfield, 2001; Penfield & Camilli, in press; Penfield & Lam, 2000; Roussos & Stout, 2004a).

8.1. Inferential Testing Versus Estimation of Effect Size

A number of authors have written about statistical testing versus estimation (Camilli & Shepard, 1994; Holland & Thayer, 1988; Kim & Cohen, 1995). Inferential test statistics are not appropriate as measures of the practical size of DIF, and they should not be used as effect sizes. Rather, parameter estimates such as the log-odds ratio or *p*-value difference should be used to express the degree of differential measurement. For example, effect sizes should be computed for comparing DIF indices across time or test administrations (Longford, Holland, & Thayer, 1993). The parallel idea exists in meta-analysis where effects sizes are used to portray the magnitude of treatment interventions. Descriptive statistics summarizing DIF across a set of items, such as central tendency or variability, should also be expressed with effect size measures. The principle here is that measures of statistical significance are not the same as indicators of *practical significance*.

8.2. Type 1 and 2 Errors

Most testing programs examine test items for group difference in measurement properties. An item is flagged when a statistically significant difference between two groups is found. However, because a statistical test is not perfect, sometimes a flagged item is a false positive and is actually measurement invariant. New testing techniques have resulted in a vast improvement in the reduction of false positive or type 1 errors (e.g., Penfield, 2001). Most problems of this kind (i.e., the actual type 1 error rate is greater than the nominal level) occur when the reference and focal groups have a large mean difference (e.g., Penny & Johnson, 1999).

An interesting trade-off in statistical analysis is the tension between type 1 and 2 errors. Type 2 errors occur when an item functions differentially, yet a statistical test fails to flag the item. A type 1 error rate (alpha level) can be set either to make fewer type 1 or fewer type 2 errors, but not both simultaneously: more type 1 errors imply fewer type 2 errors and vice versa. In most scientific work, type 1 errors are anathema, and much basic research is devoted to developing tests with accurate type 1 error levels. However, the situation in fairness work, where the "cost" of a type 2 error may be high for an examinee, is different from that of quality control in ongoing programs. In contrast, the "cost" of a type 1 error would be high to the test developer because of item development costs and rescoring procedures that may be necessitated by a deleted item. Favoring the examinee would lead to a strategy of reducing type 2 errors at the cost of more type 1 errors. Yet the trade-off should be considered in the context that flagged items are not automatically rejected; they must be reviewed for substantive interpretations of unfairness.

In a hybrid approach to signaling DIF items, Educational Testing Service uses three categories for the degree of DIF in test items (see section 7.3). Each of these categories combines a statistical test with an evaluation of effect size. This approach seems very prudent as a strategy for combining practical considerations with those of statistical hypothesis testing. It is not clear whether the utility of decision making (relative to type 1 and 2 errors) based on this approach has been compared with purely inferential testing strategies.

8.3. Item Statistical Discrimination

Wright, Mead, and Drada (1976) argued that high discrimination indices (e.g., point-biserial coefficients) for some test items could be signs of a problem with a test item. They wrote that

> high SES pupils of a given ability are more likely to be familiar with "sonata" than are low SES pupils of the same ability because of differences in exposure to this culturally biased word. . . . Typically, high SES students perform better on achievement tests . . . The greater the difference in the levels of group achievement, the more effective such culturally biased items will appear. If items are selected based on high discrimination, culturally biased items will be selected, producing tests with greater and greater bias. (p. 4)

(Note that in this context, the word *discrimination* is a technical term describing the effectiveness of an item in distinguishing those who demonstrate higher from those who demonstrate lower levels of proficiency. It has no social or legal connotations.) Masters (1993) gave a similar example involving native and nonnative speakers of a language. He

611-190-007.indd 240 7/14/2006 4:18:24 PM

argued that an item unclear to native speakers could be thought of as ideal for assessing second-language listening comprehension. One highly discriminating item was identified in which native German speakers did much better than Dutch speakers, probably because the item was based on a conversation about German politics. The contaminating influence of a second dimension can manifest itself in unusually high item discrimination (Masters, 1988), though unusually low discriminations also merit investigation. This claim is given some mathematical justification in section 8.5.2.

8.4. The Multidimensionality Hypothesis

The multidimensionality hypothesis for DIF is attributable to Hunter (1975), who illustrated, using IRT methods, that there is no conceivable way that a unidimensional test can adequately deal with group differences in the distributions of secondary abilities. He also demonstrated that, when two groups differ in average achievement level, point biserial correlations will give a false impression of bias when an item functions identically for two groups; biserial correlations had been proposed earlier as a method of DIF detection by Green and Draper (1972). Using modern measurement theory, Hunter also effectively undermined methods proposed by Angoff and Ford (1973) and Jensen (1974, 1975). Significantly, Hunter proposed multidimensional IRT models for analyzing dimensionally complex test items.

Based on the work in multidimensionality by Stout (1990), Shealy and Stout (1993b) identified three components required to produce statistical test bias: (a) the potential for bias, (b) dependence of correct item responses on secondary test factors, and (c) the test-scoring method. In this case, if two groups differ on the secondary test factors, this difference still must be transferred into a test score by a procedure for scoring. Along these lines, Gelin and Zumbo (2003) showed that an item can show DIF for some methods of scoring but not others using the same detection procedure. They concluded that DIF is a property of the item, scoring method, and purpose of the instrument because the scoring method is dependent upon the purpose of the instrument (p. 65). This underlines the interesting point that a test's purpose creates the conditions in which DIF arises.

Roussos and Stout (1996) proposed a schema for understanding how multidimensionality creates differential measurement on tests. They exhaustively considered the technical requirements for differential measurement to occur. Most importantly, they showed how test items can be combined in DIF analyses to obtain better estimates of DIF as well as more powerful tests (see also Bolt & Stout, 1996; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Nandakumar, 1993). Roussos and Stout (1996) also examined the nature of the matching (for comparability) criterion. If a test contains a number of knowledge areas or skills, examinees matched on the total score are not necessarily matched on separate dimensions that items assess (O'Neill & McPeek, 1993). Roussos and Stout (1996, p. 367) concluded that

> if the matching criterion is multidimensional, then a statistical test for DIF may reject [the null hypothesis for] a perfectly fair item simply because the examinees differ on one of the auxiliary dimensions (secondary dimensions intended to be measured by the test). Thus, with regard to Type 1 error, perhaps the most insidious cause has been the assumption of a unidimensional matching criterion when in fact the test is measuring either more than one primary dimension or a single primary dimension with several auxiliary secondary dimensions.

That is, the matching criterion, though multidimensional, may not parallel the dimensional content of a particular item. In this case, if examinees are matched on separate measures of the subskill areas, DIF may be diminished or disappear altogether.

This may be true, but it raises several issues. First, if a single score is to be used to make decisions about an examinee, then the criterion of interest is the conditioning or matching score that is most faithful to the intended construct with all of its potential multidimensional complexity (Bolt & Stout, 1996; Camilli & Shepard, 1994, Camilli, 1992). Subdividing the criterion begs the question of the proper dimensional mix of the test; in fact, this is what DIF analysis is most effective at uncovering. Second, a moderate upward bias in the type 1 error rate, as suggested above, may not be such a bad thing in detecting item unfairness.

8.5. DIF Signals: Parameters or Secondary Proficiencies?

Differential item functioning can be conceptualized either as differences in item parameters, or as differences in secondary distributions of ability (a topic similar to that of the previous section). Though the two approaches are related, there are two primary considerations for choosing between them. First, in practical applications the most simple and direct question is often "For examinees of equal proficiency, is this item more difficult for one identifiable subgroup?" The drawback of this approach is that if the answer to this question is "Yes," there is little information for assisting an explanation for *why* the item is more difficult. In the second approach, the reason for the observed difference in item difficulty is postulated to have its origin in multiple proficiencies. While a test typically measures a single dominant dimension (Reckase, 1979; Shealy & Stout, 1993b; Stout, 1987), several other secondary proficiencies may also contribute to item performance. Accordingly, observed differences in item performance for groups of equal ability may be due to group differences in the distributions of one or more secondary proficiencies; once these are identified, a means exists for understanding the nature of the DIF. The basic assumption here is that DIF originates in the knowledge and skills of examines, not the measuring properties of items.

Below several conceptual models are presented for illustrating these issues by using the parameters of IRT model. These models incorporate both item discrimination parameters (labeled α), item difficulty parameters

(labeled β), person *proficiency* (labeled θ), and a random measurement error (labeled ε). Finally, the concept of propensity (labeled z) is used to *indicate* the overall likelihood that a person receives a higher score on a test item.

8.5.1. Difference in Parameters

Differences in one or more item parameters can be expressed as

$$z_G = \alpha_G \theta - \beta_G + \varepsilon_G. \quad (31)$$

For two groups G = (r and f), if $\alpha_r = \alpha_f = \alpha$, then DIF is said to be *uniform* across θ. In this case, the expected or average DIF is

$$E[z_r - z_f | \theta] = -(\beta_r - \beta_f). \quad (32)$$

Here, DIF is expressed as the simple conditional difference between the difficulty parameters of two groups. This is because the term $\alpha\theta$ subtracts out when taking the difference, and the error terms average to zero across examinees. If the condition of equal αs is not met, then the expected difference in performance is a more complex function:

$$E[z_r - z_f | \theta] = (\alpha_r - \alpha_f)\,\theta - (\beta_r - \beta_f). \quad (33)$$

In this case, the expected conditional difference depends on where along the θ dimension the group difference is examined—this DIF is nonuniform (see Figure 7.6). Holland and Thayer (1988) showed that if items on a test followed the Rasch model, then the expected value of Mantel-Haenszel log-odds ratio for a single item having DIF is identical to the right-hand side of equation (32).

There are two general methods of estimating DIF in the nonuniform case. First, both the discrimination and difficulty parameters can be estimated using IRT techniques and then tested for group differences. The second method is to aggregate the *absolute values* of across the group distributions of θ. This can be done with either IRT techniques (e.g., Raju, 1988, 1990), observed-score techniques (e.g., Zumbo, 1999), or nonparametric techniques (e.g., Shealy & Stout, 1993a, 1993b). In all cases, construction of the conditioning variable is of paramount importance. Absent a sufficient statistic for conditioning, Zwick (1990) showed that all DIF estimates contain some degree of statistical bias.

8.5.2. Difference in Distributions

As explained above, DIF may arise due to group differences in secondary distributions of proficiency. A heuristic model for one primary and one secondary factor can be expressed for a single group as

$$\begin{aligned} z &= \alpha_1 \theta + \alpha_2 \eta - \beta + \varepsilon, \\ \eta &= \kappa + \gamma\theta + \nu. \end{aligned} \quad (34)$$

Here z is a propensity score, two proficiencies (θ *and* η), two items discriminations (α_1 *and* α_2), a single difficulty parameter β, and a random measurement error ε. Another important aspect of this model is the relationship between the two proficiencies, which is expressed as the linear regression in the second line of (34). The reduced form equation shows that for group equation G,

$$\begin{aligned} z_G &= \alpha_{G1}\theta + \alpha_{G2}(\kappa_G + \gamma_G\theta + \nu_G) - \beta_G + \varepsilon_G \\ &= (\alpha_{G1} + \alpha_{G2}\gamma_G)\theta - (\beta_G - \alpha_{G2}\kappa_G) + (\alpha_{G2}\nu_G + \varepsilon_G). \end{aligned} \quad (35)$$

If it is assumed in this case, that $\alpha_{1r} = \alpha_{1f} = \alpha_1$ and $\alpha_{2r} = \alpha_{2f} = \alpha_2$, and $\gamma_r = \gamma_f$, and it is possible to condition on the primary proficiency θ, then

$$E[z_r - z_f | \theta] = -(\beta_r - \beta_f) + \alpha_2(\kappa_r - \kappa_f). \quad (36)$$

With a strict interpretation of the multidimensionality hypothesis, *all* of the conditional expectation is due to the secondary trait, and so equation (36) can be rewritten as

$$E[z_r - z_f | \theta] = \alpha_2(\kappa_r - \kappa_f). \quad (37)$$

As above, the conditional group difference does not involve (either) proficiency, but the form of the difference indicates that the item's loading on the secondary dimension η biases its estimate of difficulty. The conditional difference (i.e., DIF) in (37) is a function of the common loading on the secondary dimension and group distributional differences in θ as expressed by different group regressions of η on θ. If the multidimensional model in equations (34) and (35) is correct, application of a unidimensional model in this case will thus confound the difference in item difficulty with the item and person parameters of the secondary distribution.

If the equality conditions required for equation (36) are not met, then the expression for the expected difference becomes unwieldy. In most scenarios where the equality assumptions are not met, the conditional difference will be a function of θ, but little else of a simple nature can be concluded. In this case, one can estimate all the parameters of the latent distributions with multidimensional IRT models with the advantage that estimation of the α_2 coefficients leads to identification of the secondary dimension (or dimensions). Once this is done, a mechanism exists for potentially connecting DIF to the educational histories and opportunities of the examinees. Through programmatic research, this may speed the development of more appropriate tests, or provide guidance for instruction. In operational applications, it is more practical for a number of reasons to estimate the average difference between the item response functions using observed score or nonparametric approaches assuming uniform DIF.

8.6. Item Difficulty Variation

Most single analyses of differential item functioning compare the item performance of one focal to one reference group, and the "group" identifier is nearly always a proxy for opportunity variables. In this manner, the notion of "group" can be recast as many levels along one or more measures of opportunity. In other words, there may exist

611-190-007.indd 242 7/14/2006 4:18:24 PM

many groups to be compared, and focal versus reference is a reduction of an opportunity continuum. For example, in many states there is tremendous variation among schools (or school districts) in resource variables such as per pupil expenditures, teacher qualifications, instructional materials, building safety, and the like. "School" provides not only an "address," but also access to many process variables collected on a regular basis.

Multilevel mixed models can be used to address the multiplicity or continuum of "groups." Accordingly, data may be analyzed in which item responses are nested within students who are in turn nested within schools. In this design, items receive a multivariate coding and are formally represented by a Rasch model embedded within the random effects (e.g., schools) model (Raudenbush & Sampson, 1999). In the embedded Rasch model, each item is represented by a difficulty parameter. Formally, this is the "fixed" part of the model, which is to say the common part of the model shared by all schools. A number of random effects are also defined within the model, and these describe between-school variation. In one approach (Camilli & Monfils, 2003; Prowker & Camilli, 2004), item difficulties can be conceptualized to have both a fixed (shared) component and a random (school specific) component. Formally, the random component describes variation in item performance across levels of an opportunity variable. For short, this variance component is labeled IDV for Item Difficulty Variation. This index describes the degree to which success on a particular item that is independent of overall proficiency varies by school. If a school does better than expected, this may reflect value added by school factors, including instruction.

Mathematically, this model for a particular item, say j, is represented by

$$f\left(n_{+sj}/n_{sj}\right) = \mu_s - \delta_{sj}, \tag{38}$$

where n_{+si} is the number of correct responses in state s and n_{sj} is the total number of responses, and $f(\bullet)$ represents the logistic link function used to linearize the relationship. In equation (38), μ_s represents the overall math proficiency of state s, and δ_{sj} represents the difficulty of item j for schools s. Two terms of the above equation can be resolved into fixed and random components:

$$\mu_s = \mu + \theta_s, \tag{39}$$

and

$$\delta_{sj} = \delta_j + \upsilon_{sj}. \tag{40}$$

In this formulation, θ and υ are random variables, and are defined as

$$\theta_s \sim N(0, \sigma_\theta^2), \tag{41}$$

and

$$\upsilon_{sj} \sim N(0, \tau_j^2). \tag{42}$$

Here, the interest is in the parameters

$$\tau_j^2 = Var(\upsilon_{sj}), \tag{43}$$

for each of the $j = 1, 2, 3, \ldots, J$ items on the test or assessment instrument. Each parameter represents how much an item's difficulty varies across schools (or other level of aggregation) independently of school θ_s and fixed item difficulty δ_j. The IDV may indicate items that are more or less instructionally sensitive, and opportunity to learn provides the motivation for examining DIF rather than unfairness.

9. RACE, ETHNICITY, OPPORTUNITY, AND EXPLANATION OF BIAS

Racial and ethnic groups are often referred to as protected groups. Analyses for detecting statistical bias are fundamentally dependent on the existence of groups, yet the definition of "group" can be elusive. If a group is defined in terms of the qualities of its individual members, yet all members within this group are not alike, then how can the meaning of differential prediction or DIF be understood? This seems to have something to do with both similarities and differences among groups, but, in turn, these qualities might depend on the purpose of defining the group. If it is assumed that examinees are alike for some purposes but not others, then what is the underlying reality to the group distinction? These questions are examined in this section.

The focus below is on the status of distinctions based on race and ethnicity, though parallel arguments could be made about other kinds of groups. In particular, the issue is the difference between individual and group interpretations of fairness. For example, one could ask whether a test using English language sentences could disadvantage an *individual* nonnative speaker of English, but on the other hand, one could ask whether *groups* of nonnative speakers are disadvantaged relative to *groups* of native speakers. Although this may seem like the same question, there is an important difference. In addition to the assumption that native and nonnative speakers are different in terms of linguistic processing, the latter question makes the additional assumption that *both* native and nonnative speakers are similar enough to be categorized separately. In other words, individuals *within* each group are assumed similar for the purpose of comparison.

9.1. RACE AND FAIRNESS

Material on race is included in this chapter because of the large differences commonly encountered in test scores among groups of different races and ethnicities, and it is important to understand the extent to which these differences are artifacts of a test rather than true proficiency. Regardless of the origin of racial differences on tests, test professionals are obliged to do everything in their power to remove sources of invalidity in the test or testing situation that contribute to these differences. Although not as frequently as in the past, many people have interpreted such differences as fixed consequences of genetic endowment. Because such interpretations are especially prone to racial stereotyping, we should seek to understand issues of human capital (e.g., ac-

611-190-007.indd 243 7/14/2006 4:18:25 PM

cess to education or medical care) as well as strive to eliminate group differences from tests that are irrelevant to the test construct. Selection bias and differential item functioning can be assessed by comparing test or item performance in different identifiable groups. Yet as suggested above, the existence of the "groups" in question is a central assumption that cannot be taken at face value. A classification of individuals exists relative to a particular purpose, and it is only this purpose that leads to identification of groups. The assumption that the classification exists with some independent reality runs the risk of stereotyping and is fraught with numerous difficulties including: the complexities of racial identity, multiracial heritage, confusion of race and ethnicity, and the potential difference between self-selected and observer classification. Below, it is argued that to understand—rather than to detect—differential test or item performance, one needs to dig deeper than commonly used group labels.

9.2. Group Membership and Social Address

Bronfenbrenner and Crouter (1983) argued that research that relates a macrosystem such as group identity to an outcome of interest employs a *social address model*. While such a model might correctly reveal a statistical connection between group (the "address") and individual outcomes, it would not clarify the processes that might explain the connection. In other words, the address tells you where you are, not how you got there.

"Social address" measurements make global assumptions about students "at the same address." Examples of social address labels are race, ethnicity, religion, political inclination, and so forth. The different routes to this address, contain the desirable or even necessary information for explaining variance in an outcome of interest. According to de Graaf (1999),

> One must look for mediating processes which link different developmental outcomes with the address label, instead of comparing people from different categories with each other, as is the case in the so called "social address research" (Bronfenbrenner, 1986). For example, stating that socio-economic-status (SES) or parental IQ affects children's cognitive achievements, does not increase in the slightest our insight into how it does this, into what kind of explanatory model is appropriate to account for those correlations. That is, correlation is not causation. (p. 72)

Reese, Balzano, Gallimore, and Goldenberg (1995) wrote that one criticism of social research has been that individuals with a common race or ethnicity have the same experience when within-group variability may be greater than between-group differences. Using the "social address" approach to group comparisons, classification into groups might be confused with a fixed biological or ethnic classification. As John Stuart Mill (1848) wrote,

> Of all the vulgar modes of escaping the consideration of the effect of social and moral influences on the mind, the most vulgar is attributing the diversities of conduct and character to inherent natural differences. (p. 319)

Lan, Brandley, Tallent-Runnels and Hsu (2002) noted that findings based on social address variables are also easily misconstrued because they sometimes may imply that there are fixed intellectual advantages or disadvantages associated with factors such as SES, ethnicity, or family composition.

There are two additional drawbacks of using labeling variables for fairness analysis. First, even the inference (statistical or otherwise) of no bias could potentially reinforce the apparent reality of the classification, as opposed to refuting its existence. Second, because many individual-level influences are submerged in a "social address," it is likely that the use of such variables will not provide a very powerful means of detecting statistical bias. Thus, the failure to reject hypotheses of differential measurement may not provide convincing support that tests or test items are, in fact, fair.

Whereas the causal examination of ecological influences on performance clearly requires more than social address variables for understanding cause and effect, it is also clear that the "addresses" are laden with social and cultural meanings that affect the lives of examinees. Thus, micro- and macro-processes can be understood in two different ways. The first is that of the analyst or scientist who seeks an understanding of social and educational processes. The second purpose is that of the institutional admissions officer who must balance the two perspectives described by Levin (2003):

> For some, fairness requires treating people as individuals, and for others, fairness requires taking into account the collective representations that matter in society. Ferdman (1997) frames this fairness debate in terms of a distinction between the "individualistic perspective" and the "group perspective." Proponents of the individualistic perspective argue that it is unfair to pay attention to ethnicity because ethnic group memberships should not influence the opportunities and outcomes of individuals in society. Proponents of the group perspective, on the other hand, argue that it is unfair not to take ethnicity into account because of the power differentials that exist between ethnic groups in society. According to this latter perspective, ignoring ethnic group membership obscures the significant ways in which these power differentials influence the opportunities and outcomes of members of different ethnic groups. (p. 8)

Both aspects of fairness must be considered in an evaluation of test bias. The scientist's approach to explaining fairness at the individual and group level should not be taken to imply that the "social labels" have no inherent meaning; indeed, stereotyping and historical discrimination are directed precisely toward the "label" rather than the individual.

10. SENSITIVITY REVIEW

The Office for Minority Education at ETS (Office for Minority Education, 1980) concluded that quantitative studies of item fairness were not likely to result in a set of practical guidelines to prevent cultural influences from interfering with test performance, and an approach coined *sensitivity*

review was recommended for developing tests that are socially balanced and evenhanded. According to Bond, Moss, and Carr (1996),

> "Sensitivity review" is a generic term for a set procedures for ensuring (1) that stimulus materials used in assessment reflect the diversity in our society and the diversity of contributions to our culture, and (2) that the assessment stimuli are free of wording and/or situations that are sexist, ethnically insensitive, stereotypic, or otherwise offensive to subgroups of the population. (p. 121)

A panel of trained reviewers is required for examining each item on a test or assessment. Panelists operate with the principle that all students should be treated equitably, and should have a common understanding of the questions and tasks. All items on a test should be reviewed at least once for this purpose, and this process has become especially important for performance items (Bond et al., 1996).

Procedures for creating culturally sensitive tests may seem like unnecessary "add-ons," thus spurring the complaint of political correctness. However, the motive here is to follow the normal professional standards of avoiding irrelevant difficulty, which includes distracting or offensive language, because it is important for test developers to create the least stressful environment possible for test takers. Sensitivity review is a method of procedural due process in establishing test validity; as such, it should be documented in rich detail. Public scrutiny is especially important in preventing both overzealous and lax evaluations.

Below, I draw heavily from documents produced at the Educational Testing Service (ETS) and ACT, Inc. In particular, ETS (2003) provides much valuable detail on constructions to avoid in test item development. These formal documents have an established historical use, are easily available, and are free of charge.

10.1. Panel Formation

A panel of sufficient size (a minimum of 5–10) is usually required for large-scale assessments, and it should be clear *how* panelists were selected. The specific procedures for selecting panelists can vary according to circumstances and may involve both nomination and self-selection. In any event, it is incumbent on the test developer or program personnel to provide public access to descriptions of panel membership (e.g., gender, race, position). Members should also have professional or instructional experience in the subject matter of the test. For example, teachers at the secondary and postsecondary level are prime candidates for sensitivity review of college entrance examinations. Panelists must be able to understand and represent a culturally and ethnically diverse range of perspectives. This is facilitated by a panel that is itself diverse in terms of factors such as race, ethnicity, gender, and geography. As a rule, test authors and item writers cannot provide a sensitivity review of their own work: sensitivity review is an *independent* review to detect unintentional language and biases in the test material.

10.2. Sensitivity Training Procedures

Training should be based on a written policy expressing commitment to the objectives of fair testing, and a set of guidelines compiled in a formal document (Office of Minority Education, 1980). Panelists should be familiar with both the test specifications and sensitivity guidelines. As noted by Ramsey (1993) training is initiated by presenting to panelists a sample of test items and then asking for a set of judgments. These questions include whether there is a problem with an item, to which guideline the problem relates, potential item revisions, and whether modification or deletion of an item should be mandatory. The goal of training is for reviewers to arrive at a consensus on these judgments as they proceed through the set of sample items. Thus, sensitivity training corresponds to procedures used in training raters to score open response items (Bond et al., 1996; Ramsey, 1993) or in setting standards (Camilli, Cizek, & Lugg, 2001; Raymond & Reid, 2001). A typical training session might range from a half to a full day.

10.3. Training Criteria

Ramsey (1993) listed six criteria used at Educational Testing Service: stereotypes, underlying assumptions, controversial material, elitism and ethnocentricity, balance, and examinee perspective. In a more recent document (ETS, 2002), the first four areas above are listed (with "controversial" changed to "inflammatory"); balance and examinee perspective have been dropped or combined into other areas; and the categories "tone" and "inappropriate terminology" have been added. A number of helpful examples are given in ETS (2002) and Office of Minority Education (1980). Other publications (ACT, 2003, 2004) list five general considerations: offensiveness, fair portrayal, diversity and balance, fairness in language, and curriculum-based content and skills. The last consideration specifies that vocabulary, concepts, and experiences (required for understanding test items) should be appropriate for all examinee groups.

There are several general themes within these and other guideline documents. Among these is the principle that racial and ethnic categories have preferred terms. For example, the labels Black American or African American are preferred to earlier terminology, and specific descriptions of ethnicity, such as tribal names of Native Americans, should be used when applicable. Some ethnic descriptions may be preferred to others; for example, the terms *Latina* and *Latino* are more consensual than *Chicana* and *Chicano* as terms for Puerto Rican or Mexican Americans.

The possibility of unintentional stereotyping is a concern in the item writing stage of test development. This concern can include, but is not limited to a number of stereotype categories including, cultural, regional, occupational, religious, and Eurocentric. Implications implying superiority with respect to group status may be difficult to identify; and especially items using the word "minority" may have a subtle implication that minority status is akin to a fixed trait. Similarly, ethnocentrism may be subtle as illustrated in following passage from a test item (Ramsey, 1993, p. 383):

611-190-007.indd 245 7/14/2006 4:18:26 PM

> The Inuvialuit, what the Eskimos of Canada's Arctic prefer to be called, live in the treeless tundra around the Bering Sea.

The problem here is the implication that the real name of the people is Eskimos, that is, the *real* name is determined from the perspective of a different culture.

Condescending material and inflammatory material is less subtle. Reference to a woman as the "lady lawyer" (Office of Minority Affairs, 1980, p. 49), or gratuitous reference to controversial material such as prayer in school, is not likely to be included during item development. However, controversial material may be both appropriate and necessary given the content specifications of a test. An assessment developed for a social studies chapter on *Roe v. Wade* (1973) will necessarily involve contentious or disturbing topics as will the study of euthanasia in Nazi Germany; yet such material would be dubious on a typical test of reading comprehension because it may introduce distractions influencing student performance. Preparing tests for international populations requires additional safeguards.

According to ETS (2003) guidelines, when issues of gender orientation are construct relevant:

> The words *bisexual*, *gay*, *lesbian*, and *transgendered* are all acceptable. Because some people assume that gay refers only to men, use *gay* or *gay people* only when prior reference has specified the gender composition of this term. (p. 21)

References to gender orientation are not consensual (e.g., queer versus gay) and the specific assessment contexts must be carefully considered. Other criteria for sensitivity review include religious beliefs, English as a second language, disability, socio-economic status, and violence. Guidelines should also pertain to gender references in language, and require revision of references to the generic "he" to more gender-neutral constructions (American Psychological Association, 1977).

One last area in this brief sketch of sensitivity review is representational balance. Test items, passages, tasks and so forth should represent population diversity including cultural references, gender roles, disability, and ethnicity. Of course, this does not mean that every test should be perfectly balanced. Rather, the goal is to achieve a reasonable representation of the appropriate populations (American Psychological Association, 1977), to respect the beliefs and experiences of all test takers, and to provide a minimum of distracting content.

10.4. Panel Operation

Formal review can begin with the design of test specifications prior to first drafts of test items, but informal review may occur at any stage of item development. Parallel issues exist in the scoring of examinations, directions to teachers and students, and score reporting. Sensitivity review can be broadened to include scoring procedures (e.g., instructions to scorers) and materials (e.g., scoring rubrics) as well as materials and procedures for reporting test scores. A more detailed description of the logistics of this process is given by Ramsey (1993).

The description in ACT (2003) suggests three stages for a typical fairness review: (1) before materials are pretested, (2) concurrently with pretest item statistics, and (3) before operational forms are administered. Different reviewers at each stage may provide an additional safeguard. In the first stage, reviewers receive packets of items, including directions and prompts. Based on the reviewers' feedback, testing staff may change, modify, or delete material. In the second stage, items are reviewed with quantitative pretest information including DIF statistics. The latter are used to aid identification of items in which irrelevant factors associated with group (e.g., gender or race) contribute to item difficulty. In the third stage, intact test forms are examined. Given that items have already been reviewed once and possibly twice, there is less flexibility in making modifications on the basis of review information in this latter stage.

Fairness reviews can be conducted by mail out for document examination, and by conference or teleconference for deliberation and exchanging views (ACT, 2003). Reviewers typically examine and annotate materials prior to participating in conferences of 1–2 hours at the first and third stages. The second stage review is more likely to be conducted by testing staff (though not item development staff) rather than independently chosen reviewers or consultants. If testing materials are reviewed off-site, it is common that confidentiality agreements are necessary.

11. FAIRNESS IN CLASSROOM ASSESSMENT

In general, the equity concerns addressed by sensitivity review extend beyond the purview of formal assessment. While the technical standards and concepts of validity and fairness are most often applied to standardized tests administered annually, a parallel set of concerns exists for classroom testing, including paper and pencil tests, performance assessments, authentic assessments, and even observational measures of learning. This section concerns fairness issues with *formative* and *summative* classroom evaluation. Shepard (in the chapter on classroom assessment) defines formative assessment as "assessment carried out during the instructional process for the purpose of improving teaching or learning," while summative assessment "refers to the assessments carried out at the end of an instructional unit or course of study for the purpose of giving grades or otherwise certifying student proficiency." Any assessment that eventually affects a grade can be viewed as summative, and therefore many assessments, including standardized tests, can share both summative and formative purposes (Brookhart, 2003, 2004).

A classroom performance assessment by its very nature is more closely connected to instruction than a standardized test, and is typically given to small numbers of students. As a rule, such tests have a short life cycle, and though it would be theoretically possible to examine differential measurement, such an effort would be impractical. Other types of statistical analysis are also of limited application. In the

611-190-007.indd 246 7/14/2006 4:18:26 PM

short run, then, different criteria must be brought to bear for insuring fairness in assessment.

Beyond traditional reliability, a number of practices can enhance the fairness of classroom assessments, both in terms of obtaining an accurate evaluation, and the perception of fairness on the part of students. *The kinds of practices and activities that insure testing is fair are also those that provide a solid foundation for attaining learning goals.* Thus, fairness is inherent in, rather than external to, effective classroom instruction. Brookhart (2004) noted the central role of construct validity:

> Particularly important for evaluating the validity of a classroom assessment is defining the construct in its instructional and contextual sense. Is writing part of the skill or irrelevant? Did we discuss this in class? Is this problem the same one as the students have already worked (in which case it measures recall)? Is this problem similar to the ones the students have already worked (in which case it measures transfer)? A close transfer or a stretch? How far did you want them to go? (p. 10)

Given that a clear purpose for the assessment is defined, additional practices are generally recognized (e.g., Brookhart, 2004; Shepard, 2000; Stiggins, 2002). A number of these are briefly examined below. A theme that runs through each of these areas is that all children are not the same. Equity and fairness are insured not by treating all children identically, but by differentiating among children to provide them the most effective opportunities to learn. The chapter by Shepard gives a broader overview of classroom assessment; here the focus is only on fairness issues.

11.1. Clear and Reasonable Assessment Criteria

Students should have an understanding of the content coverage of a test, and all the material on the test is relevant to the course's learning objectives. Students should also understand the process for scoring a test, and for open-response items, this requires a working knowledge of scoring rubrics. Shepard (2001) noted the dual purpose of communicating to students the standards by which their work is judged:

> Although access to evaluation criteria satisfies a basic fairness criterion (we should know the rules for how our work will be judged), the more important reasons for helping students develop an understanding of standards in each of the disciplines are to directly improve learning and to develop metacognitive knowledge for monitoring one's own efforts. These cognitive and metacognitive purposes for teaching students explicitly about criteria then speak to a different sense of fairness than merely being even-handed in evaluating students, that is, they provide students with the opportunity to get good at what it is that the standards require. (p. 1093)

The use of rubrics that define the general characteristics of good work should make assessments fairer in the long run because they help students to visualize their target (Brookhart, personal communication, September 6, 2004). Students should also be exposed to task-specific rubrics, though obviously not prior to a summative assessment activity.

Students should have equivalent understandings of the grading criteria, and according to these criteria, teachers should grade consistently. For example, criteria should remain consistent from the first to the last test or product scored on a particular occasion, and a rubric should be used long enough for students to acquire familiarity. A rational marking system might provide numerical standards for arriving at a grade, or combine a numerical system with judgment in a logical manner. The latter could be implemented, for example, by first establishing a numerical system for marking, and then adjusting grades based on judgment. This might benefit all students if adjustments are based on corroborating evidence from other instances of the same academic work, but absent a systematic and empirical procedure, an "intuitive" adjustment may cause more harm than good. Test scores should be weighted to reflect content area emphases or importance of a particular mark. For example, if a student's grades consist of numerous grades of A and one B+, and the final grade is B+, this provides an inconsistent message about a teacher's standards. Finally, explicit policies are required regarding makeup performances, for transfer students, for students working on group projects, and so on.

11.2. Equity in Assessment and Instruction

Effective classroom assessment supports student learning, and effective assessments, in turn, encourage students to focus on the task rather than their own level of competence *per se* (Black & Wiliam, 1998). Gipps (1999) elaborated:

> Children's evaluations of their ability and feelings toward themselves are more negative when the classroom climate is focused on winning, outperforming one another, or surpassing some normative standard than when children focus on trying hard, improving their performance, or just participating. (p. 383)

Equity is well served by maintaining focus on the learning culture (Shepard, 2000). As Moss (2003, p. 19) wrote, "validity in classroom assessment—where the focus is on enhancing students' learning—is primarily about consequences." It follows that the validity of assessments can be judged in part by their consistency with task-oriented learning activities presented at the appropriate level of difficulty. At its best, assessment in the classroom is more like instruction than testing.

11.3. Opportunity to Learn

The issue of equity raises a number of questions about opportunity to learn. Students should have an opportunity to learn all of the material on a test. In addition, they should understand what opportunity to learn means, and how it can be recognized. Testing for skills and knowledge not taught and practiced in class is problematic unless explicit directions are given regarding learning outside the classroom. Teachers should also consider whether students have equal access to knowledge and instruction. For example, with take-home assignments that are graded,

what role should parents have in assisting their children, and how should teachers compensate for such differential access to knowledge?

Research in educational psychology has demonstrated that children and adults learn by incorporating new information and concepts into extant schema. One goal of classroom assessment is to "locate" the presenting proficiencies of students in order to guide instruction along the contours of their strengths and weaknesses. Students who enter at a more advanced level have more places to attach new meaning to old, and it is unreasonable to define fairness as the elimination of individual differences. Teachers can, however, unpack their lesson objectives to see if there is anything that, if not addressed, sets certain students up to fail or otherwise miss significant opportunities to learn. In this regard, it is important for teachers to experience (through initial training or professional development) what students are able to accomplish with help designed on the basis of formative assessment.

11.4. Sensitivity and Construction of Assessments

Test content should be free of cultural, ethnic, racial, religious, and gender stereotypes. It is important, especially with young test takers, to avoid stereotypical material since these children are, to a large extent, learning about themselves through the process of testing. This may require teachers to become familiar with the backgrounds and beliefs of students and their parents. Pellegrino, Chudowsky, & Glaser (2001) warned that

> Apart from the danger of a teacher's personal bias, possibly unconscious, against any particular individual or group, there is also the danger of a teacher's subscribing to the belief that learning ability or intelligence is fixed. Teachers holding such a belief may make self-confirming assumptions that certain children will never be able to learn, and may misinterpret or ignore assessment evidence to the contrary. (p. 240)

Also, familiarity with the prior knowledge of students can help to insure that test items and tasks are developed at an appropriate level of difficulty. Distributions of scores or marks with notable floor and ceiling effects might indicate inappropriate difficulty, lack of clarity, or insufficient instruction.

11.5. Multiple Measures

Assessment tasks should be designed to reflect accurately what students know and can do. Yet it is unlikely that all learners can demonstrate what they know in the same way. According to Shepard (2001), students should be given an accessible opportunity to show their knowledge and this could be provided, for example, by an oral presentation rather a written exam, writing about a familiar topic, or providing translations. These examples illustrate that multiple opportunities should be provided for a student to demonstrate competence; and these opportunities should include alternative assessment formats.

Especially for classroom assessment, the phrase "multiple measures" does not mean multiple opportunities to pass the same test, or even equivalent tests. As in most aspects of fairness in classroom assessment, genuinely alternative measures have a dual role. Not only do multiple measures provide an equitable assessment, they also aid the process of learning. They can facilitate knowledge transfer as well as the diagnosis and remediation of language deficits.

11.6. Modeling Fairness

Because the instructor (or teacher or professor) is the grading authority in the classroom, grading can be a type of modeling of both learning and self-assessment of students. This imparts a serious responsibility to the instructor. In an overt approach to grading, the criteria are explicitly communicated, and students are oriented toward learning rather than social evaluation of themselves or other students. Students in this environment will likely discover that learning depends on effort. When internalized, this becomes a model process for the students' participation in their communities and a larger democratic culture.

12. A BRIEF HISTORY OF COLLEGE ENTRANCE EXAMINATIONS

The measurement topics covered above have all evolved historically, spanning, for the most part, the latter half of the 20th century. In this section, a broader perspective is provided by considering testing issues of the 19th to mid 20th century. Trevelyan and Northcote (as cited in Gipps, 1999, p. 357) wrote with respect to the civil services,

> We are of the opinion that this examination should be in all cases a competing literary examination. This ought not to exclude previous inquiry into the age, health, and moral fitness of candidates. Where character and bodily activity are chiefly required, more, comparatively, will depend upon the testimony of those to whom the candidate is well known; but the selection from among the candidates who have satisfied these preliminary inquiries should still be made by a competing examination. This may be conducted as to test the intelligence, as well as the mere attainments of the candidates. We see no other mode by which (in the case of the inferior no less than superior offices) the double object can be obtained of selecting the fittest person, and of avoiding the evils of patronage.

In this remarkable quotation from 1853, a number of common 20th and 21st century ideas are present. Testing is recognized as a tool for social change, and the distinction is made between intelligence and achievement. The authors presumed that candidates could be ordered along a dimension of "fitness," and also recognized that different types of evidence may be required.

By 1869, Sir Francis Galton had published several articles and a book, *Hereditary Genius*, in which he began to augment informal notions of intelligence. In particular, he suggested that the human traits of "great ability" were inherited, a conclusion that was inspired by the work of

611-190-007.indd 248 7/14/2006 4:18:26 PM

his cousin Charles Darwin (Porter, 1986). Based on this idea, Galton proposed the *eugenics* thesis: selective breeding could improve the human species as much as any other (Zenderland, 1998). A number of Americans were deeply influenced by this work including Henry H. Goddard, who translated Alfred Binet's original IQ scales into English; Lewis M. Terman of Stanford University, who adopted Stern's ratio conception of intelligence and coined the shorthand term IQ; and E. L. Thorndike, a pioneer in psychometrics. Intelligence testing in America began in earnest during World War I, when Robert M. Yerkes led a team of social scientists, including Terman and Goddard, Walter V. Bingham, and Carl C. Brigham in adapting Terman's Stanford-Binet intelligence test for group administration to army recruits. This test, the Army Alpha, was used to collect large amounts of data, but the project generated controversy and had little, if any effect on selection and placement during World War I (Gould, 1996; Hartigan & Wigdor, 1989). A number of these and other intelligence theorists and psychometricians were involved in the eugenics movements of the early 20th century (Zenderland, 1998; Lombardo, 2002, 2003); however, an examination of this thread is related beyond the scope of the present chapter.

American and English universities and institutions began using selection and placement examinations as early as the 1850s, and intelligence tests were also thought to hold great promise in the upper reaches of society as tools for selection and placement in higher education. For example, Lazerson (2001, p. 386) noted "by the end of the 1920 academic year, over 200 colleges and universities had given intelligence tests." In the following two sections (12.1 and 12.2), a brief sketch is presented concerning how intelligence testing and the new psychometrics developed relative to two major college entrance examinations. In section 12.3, these developments are linked to current issues in test fairness.

12.1. The SAT

After World War I, Brigham (whose mentor was Yerkes) became a professor at Princeton University and began administering his own version of the Army Alpha—the Princeton Psychological Examination—to Princeton freshmen. Shortly thereafter, he chaired a committee of experts for the College Entrance Examination Board (initially with Yerkes and Henry T. Moore, chair of Dartmouth's psychology department), which was empanelled for recommending a new college admissions test (Hubin, 1988). This test, the Scholastic Aptitude Test (SAT) was completed in 1925, and was given to high school students for the first time the following year. Four of the nine item types on the new tests came from existing psychological tests at Dartmouth, Smith, and Yale; the other five came from Brigham's Princeton test (Hubin, 1988).

The SAT was created to test more integrated and cross-subject thinking as well as to standardize the admission process. Lazerson (2001) noted that the College Board's 1914 Annual Report called for an examination

> so framed as to give the candidate the opportunity to exhibit his power to think independently and to compare or correlate different parts of the field. The ability to reproduce with more or less fidelity the material presented on the pages of a text book would be considered as of secondary importance. (College Board, 1914, pp. 12–13)

At the time, college entrance tests varied widely among institutions of higher education, and individual tests tended to concentrate on factual content. To some degree, this 1914 statement from the Board reflects the mounting pressure to respond to new psychometric developments. Yet the Board was constrained by its "traditional nemesis," the charge that it was attempting to control the college preparatory curriculum (Lazerson, 2001, p. 391). The new SAT was thus crafted to navigate both demands. Testing scholastic "aptitude" tapped critical thinking, yet it did so with test items that for the most part avoided specific curricular knowledge.

In the early 20th century, "merit" in the college admission process, especially at Ivy League schools, was operationally assessed primarily in terms of family privilege, attendance at a small group of eastern preparatory schools, and a brief assessment of moral character (Wechsler, 2001). Shortly after James Bryant Conant was appointed president of Harvard University in 1933, he charged two associate deans, Wilbur Bender and Henry Chauncey, to establish an ambitious new academic scholarship program for students with limited financial resources. To select students for this program, Bender and Chauncey proposed the SAT to Conant as an accurate measure of *intelligence*—a condition upon which Conant had insisted (Lemann, 1999). By 1941, the SAT was required for all Harvard applicants. In the 1950s, the College Board had 300 member institutions administering the SAT as well as a number of nonmembers (Lemann, 1999). Conant's proximal goal was to recruit talented students from a wider geographic area; his broader and enduring social goal was to break up a system of influence, money, and privilege by a fair selection process. Intelligence testing for access to education, as opposed to traditional notions of merit, seemed to provide an objective tool for this purpose.

From the very beginning of the SAT, both Brigham (Hubin, 1988) and the Board (Lazerson, 2001) sought to downplay the connection between intelligence and aptitude: the term *scholastic aptitude* according to this argument makes no assumptions other than a positive correlation with subsequent academic performance:

> The term "scholastic aptitude" makes no stronger claim for such tests than that there is a tendency for individual differences in scores in these tests to be associated positively with individual differences in subsequent academic attainment. (Brigham, n.d., p. 1)

Conant later elaborated:

> as originally developed and used by many educators in the 1920s and 1930s, intelligence tests were thought of as measuring the inherent or genetic qualities of the individual. The evidence at first available seemed to indicate that the chances of a single individual's I.Q. changing over the years were slight. Today, however, when we tend to think of paper-and-pencil intelligence tests, at least in the higher grades, as measuring a type of scholastic aptitude, we are

> well aware that we are measuring an aptitude which in part has been developed in the school. The difference between an I.Q. test and a good achievement test is one of degree not kind. Understood in this sense and with evidence accumulating that an individual's aptitude score may change during his school years, there is nothing deterministic about the use of the various forms of intelligence or aptitude tests which are on the market. If they are understood only as giving a prediction of probability of academic success in subsequent schoolwork, they are no more and no less influenced by home or other environmental factors than are the marks for schoolwork given by a conscientious teacher. (Conant, 1961, pp. 13–14)

These views are consistent with other early views of aptitude (e.g., Chauncey & Frederiksen, 1951, p. 89). Zwick (2002, p. 33) provided additional details of this development.

Despite this agnostic position on the meaning of scholastic aptitude, the SAT from its inception began to evolve. While retaining some core elements from its IQ ancestry (analogies, but not antonyms), the reasoning test (now the SAT I) was redesigned for 1994, according to Lawrence, Rigol, Van Essen, and Jackson (2003), "to reflect contemporary secondary school curriculum and reinforce sound educational standards and practices" (pp. 10-11). In 2005, the SAT I was further revised: the verbal section was retooled to assess critical reading, analogies were eliminated, and a new mathematics section added items from third-year college-preparatory mathematics.

12.2. The ACT

E. F. Lindquist, as a new assistant professor of education at the University of Iowa, founded a program in 1928 called the Iowa Academic Meet, dubbed the Brain Derby by the local press (Lindquist, 1970), which was a statewide scholastic contest designed to identify academically talented teens. Lindquist soon discerned "too much emphasis on the competitive features" of this process (Lindquist, 1970, p. 9), and became more interested in a test that would provide guidance in educating a broader population of students. In the 1930s, Lindquist developed the Iowa Every Pupil Achievement Tests (IEPT), the Iowa Tests of Basic Skills (ITBS), and the Iowa Tests of Educational Development (ITED). By the 1940s, Lindquist directed a state assessment program that

> had several remarkable features: every school in the State could participate on a voluntary basis; every pupil in participating schools was tested in key subjects; new editions of the achievement tests were published annually; and procedures for administering and scoring tests were highly structured. (Office of Technology Assessment, 1992, p. 122)

With Phillip Rulon of Harvard, Lindquist in the 1950s designed an electronic scoring machine. The first "Iowa machine" went into production in 1955, and by 1957, it is now clear that the electronic scoring machine reshaped the landscape of educational testing (Office of Technology Assessment, 1992).

Lindquist was also deeply involved with the admissions practices of colleges as a member of the SAT standing committee of College Board (Lindquist, 1970). He advocated expanding college admissions, and sought to develop a test of broad competencies for facilitating selection, and for assisting placement and guidance after selection. Lindquist (1951) had expressed skepticism regarding the use of aptitude tests for this purpose, though an intelligence test (supplied by Thorndike) was used experimentally in the Iowa Testing Program in 1934. While such a test might predict a criterion outcome, Lindquist argued that they were of little use for other educational purposes. What was needed was a test of competencies representing the same kinds of reasoning and problem-solving tasks required in high school and in college:

> the most important consideration is that the test questions require the examinee to do the same things, *however complex*, that he is required to do in the criterion situations. (Lindquist, p. 154, original emphasis)

Lindquist (1951) more ambitiously intended to create an admission test that would stimulate curricular reform; he was highly critical of the traditional high school curricula of his era as well as the extant subject-specific tests used for college admission. Though Lindquist apparently had proposed a new admissions test in which critical thinking was combined with subject matter content to the College Board in 1958 (Coulehan, 2004), he and Theodore McCarrel cofounded the American College Testing (now the initialism ACT) Program to develop a new admission test that soon became the main competitor to the SAT. The ACT test was designed to serve the needs of large state universities as well as state, municipal, and junior colleges rather than elite east-coast institutions.

As Coulehan (2004) observed, "Lindquist's greatest concern was," according to Ralph Tyler, "devising tests to gauge the educational development of each student for purposes of guidance and counseling as well as for college admission and placement." He wanted to design an achievement-oriented admissions test that would provide diagnostic information in the form of four subtest scores of the ITED (Peterson, 1983). After World War II, the Test of General Educational Development (GED) was developed to help youths and adults, especially those who were returning from the war, to demonstrate knowledge for which they would receive academic credit or a high school equivalency diploma. It is not surprising that the GED was also adapted from the ITED with substantial input from Lindquist (Peterson, 1983).

Recent ACT assessments are based on periodic surveys (e.g., in 1998–1999, and most recently in 2002–2003) of state education practices including examination of standards documents, survey of educators, and consultation with content area experts. Based on this information, the four curriculum-based tests are revised and updated (English, Mathematics, Reading, and Science Reasoning). According to current documentation (ACT, 2002), the ACT can be used for a number of purposes including advising and counseling at the high-school level; admissions, recruitment, and course planning and placement at the collegiate level; and scholarship and recognition programs. The stated philosophy of the test remains consistent with Lindquist's view "that the best way to measure students' readiness for

611-190-007.indd 250 7/14/2006 4:18:27 PM

postsecondary education is to measure as directly as possible the knowledge and skills students will need to perform college-level work" (ACT, 2002, p. 1). An optional writing test was added to the ACT Assessment in 2005.

12.3. Fairness Issues

Intelligence testing and its aptitude incarnation were conceived as indicators of merit for access to education in the early 1900s. The rationale for aptitude as a democratizing criterion, however, was greatly obscured by the subsequent claims that aptitude was nothing other than a label for a test score that correlates with educational attainment. Indeed, the developmental histories of the SAT and ACT reveal an enduring ambivalence regarding how merit should be defined and measured. According to Lazerson (2001), this ambivalence is expressed as a

> debate that became one of the most intense, fundamental, and divisive of the twentieth century: whether to measure what students knew based on what they were taught or to measure what students were capable of learning. (p. 385)

What examinees know in any given situation is a mixture of educational and informal learning experiences that is strongly affected by access and opportunity to learn. Rather than a simple "individual fair" proposition, the argument for measuring aptitude reflects a concern for recognizing talented students who are highly capable of learning (individual merit) *and* an historical efficiency preference (social or institutional benefit). Yet merit based on potential rather that actual criterion performance is deeply problematic from a philosophic point of view (Rawls, 1971). The SAT and ACT embody these different perspectives to some degree, though they are becoming more similar.

In many institutions of higher education, this dilemma is currently and sensibly resolved by allowing candidates to submit either SAT or ACT scores, to consider subject matter assessments such as the SAT II tests, and to allow for additional student background information (as originally recommended by Brigham) including interviews and high school grade point average. As suggested by Zwick (2002), however, it is not likely that more achievement-oriented selection strategies will substantially alter access to more elite institutions of higher education. Both aptitude and achievement tests reflect the lengthy educational histories of examinees, and equity issues must be addressed either by long-term infrastructure changes, or by affirmative action (as in *Grutter v. Bollinger*).

Other institutional consequences of the intelligence-testing legacy should be considered. Popham (2001, p. 46) warned that IQ-like test items, "are not suitable for evaluating schools." Tests composed of such items are consistent "with assumptions about knowing and learning that existed within the behaviorist perspective" in which knowledge is attained as a set of component skills, without regard to deeper underlying structures or representations (Pellegrino et al., 2001, p. 61). To the degree that a test is constructed with a similar logic, its items are independent samples of these knowledge "bits." Mastery is then defined in a way that is insensitive to effective instruction, on the one hand, and highly susceptible to influences on learning that are weakly related to formal instruction, on the other. As recognized by Lemann (2004, p. 14) "tests don't exist in a social vacuum. The way they are used embodies ideas about how a society should work." Tests have symbolic value regarding the importance of their content, and who can learn the content. Aptitude-dominant tests send weak signals regarding content, and potentially misleading signals regarding who can achieve standards, while achievement-dominant tests may provide clearer messages about the responsibilities of schools in preparing students for college selection and college experience.

13. CONCLUSIONS

Concerns about fairness arise from the intended and unintended consequences of testing. Fairness is thus not a property of a test *per se*, and for this reason, investigations of fairness are framed by test use. The clarity of both questions and answers in these investigations is promoted by involving key stakeholders including test users, test takers, developers and contractors, and measurement scholars. Cole and Moss (1989) recognized that

> Reponses to questions of whether a test *should* be used for a particular purpose and whether that purpose should be served . . . are the right and responsibility of all persons affected by test use. This includes test takers and others to whom the consequences of testing are of concern, as well as the measurement profession. (p. 207)

As Linn (1989) suggested, other influential groups have also increasingly begun to appear on the scene including judges, legislators, and administrative agencies.

Constructs are created in (though not by) a social context. In my opinion, there is no single "objective" point of view either for delimiting the construct, or for sorting out intended and unintended consequences. Measurement professionals certainly have a central role to play in any discussion of fairness, but must recognize other important actors. Gardener (1999) asked "Who owns intelligence?" Similarly, one could ask "Who owns developed ability?" or "Who owns achievement?" For this reason, Gipps (1999) argued that rather than analysis by specialists,

> The best defense against inequitable assessment is openness. Openness about design, constructs, and scoring will bring out into the open the values and biases of the test design process, offer an opportunity for debate about cultural and social influences, and open up the relationship between the assessor and learner. (p. 385)

But even the above questions do not go far enough because the central theme of test fairness concerns the match between a test's measurement properties and the purposes and goals for which the test is used. Indeed, an aptitude test may well suit an institution's mission, but a rationale should be provided and defended, rather than presumed. Given a sensible and clear institutional mission, the goal is to understand how decisions can be informed by the full range of evidence regarding an individual's qualifications. (With some modification, this goal is relevant to

611-190-007.indd 251 7/14/2006 4:18:27 PM

classroom practices as well.) An adequate analysis of test fairness may involve examining assumptions underlying the test construct as well as the purpose of the test, the consequences of its use, and the responsibilities of test developers and examinees. The latter are further elaborated in the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) and the *Code of Professional Responsibilities in Educational Measurement* (NCME Ad Hoc Committee on the Development of a Code of Ethics, 1995). In terms of contemporary issues, resolving fairness problems also involves closing the achievement gap, providing opportunity to perform, deterring the misuse of tests, and accommodating individual differences (Cole & Zieky, 2001).

There are now sophisticated methodologies for examining test fairness issues, and these have already helped to distinguish bias in measurement and predictive processes from fairness in test use. With this set of analytic tools, psychologists and measurement experts will continue to play key roles in developing and validating tests. Yet selection, placement, promotion, and certification are mechanisms of social and educational management (and control) that reflect societal tensions—as is the case with recent court decisions—and will continue to do so in the 21st century struggle to reconcile efficient selection with a modern vision of inclusion. Linn (1989, p. 6) recognized that "clarity in definitions and evidence regarding the comparability of prediction systems cannot be expected to resolve the underlying value conflicts."

Test developers do not operate independently of the social context in providing tools for teachers and education managers, just as teachers do not operate independently of prevailing attitudes and beliefs when giving a classroom assessment. Consequently, test developers have the challenging responsibility to construct sound tests, but also to inform clients fairly regarding the purposes, interpretations, and uses of ensuing test scores. In turn, tests users have the significant responsibility to reconcile the choice of test, as well as the potential consequences of its use, with institutional and social goals.

REFERENCES

ACT, Inc. (2002). *Your guide to the ACT assessment.* Iowa City, IA: ACT.

ACT, Inc. (2003). *Consultant's guide for the fairness review of the ACT EPAS tests.* Iowa City, IA: ACT.

ACT, Inc. (2004). *Fairness report for the ACT assessment tests.* Iowa City, IA: ACT.

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: Wiley.

Alexander v. Sandoval, 121 S. Ct. 1511 (2001).

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

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

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 107–116.

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

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*, 7–74.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23,* 67–95.

Bond, L. (1980). [Book review of *Bias in mental testing*]. *Applied Psychological Measurement, 3,* 406–410.

Bond, L. (1981). Bias in mental tests. *New Directions for Testing and Measurement: Issues in Testing—Coaching, Disclosure and Ethnic Bias, 11,* 55–77.

Bond, L. (1994). Comments on the O'Neill and McPeek paper. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277–279). Hillsdale, NJ: Lawrence Erlbaum.

Bond, L., Moss, P., & Carr, P. (1996.) Fairness in large-scale performance assessment. In G. W. Phillips & A. Goldstein (Eds.), *Technical issues in large-scale performance assessment* (pp. 117–140). Washington, DC: National Center for Education Statistics.

Braun, H. (2006). Empirical Bayes. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Complementary methods in education research* (pp. 243–258). Washington, DC: American Educational Research Association.

Brigham, C. C. (n.d.). *Scholastic aptitude tests: A manual for the use of schools.* Prepared by the College Entrance Examination Board. Document is housed in Educational Testing Services Archives, Princeton, NJ.

Bronfenbrenner, U., & Crouter, A. C. (1983). The evolution of environmental models in developmental research. In P. H. Mussen (Series Ed.) & W. Kessen (Vol. Ed.), *Handbook of child psychology: Vol. 1. History, theories, and methods* (4th ed., pp. 357–413). New York: Wiley.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*, 5–12.

Brookhart, S. M. (2004). *Grading.* Upper Saddle River, NJ: Pearson Education.

Brown v. Board of Education, 347 U.S. 483 (1954).

Burton, E., & Burton, N. W. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer H. (Eds.), *Differential item functioning* (pp. 331–335). Hillsdale, NJ: Erlbaum.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16,* 129–147.

Camilli, G., Cizek, G. J., & Lugg, C. A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 445–476). Mahwah, NJ: Lawrence Erlbaum.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous items. *Journal of Behavioral and Educational Statistics, 4,* 323–341.

Camilli, G., & Monfils, L. (2003, April). *Item difficulty variation (IDV) approach to school assessment.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34*, 123–139.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the items in a test*

(Research and Development Reports RDR-64-5 No. 9, College Entrance Examination Board; also Research Bulletin RB-64-61. Princeton, NJ: Educational Testing Service.

Chauncey, H., & Frederiksen, N. (1951). The functions of measurement in educational placement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 85–116). Washington, DC: American Council on Education.

Chung-Yan, G. A., & Cronshaw, S. F. (2002). A critical reexamination and analysis of cognitive ability tests using the Thorndike model of fairness. *Journal of Occupational and Organizational Psychology, 75,* 489–509.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17,* 31–44.

Clauser, B., Mazor, K., & Hambleton, R. (1993). The effects of purification for the matching criterion on the identification of DIF using the MH procedure. *Applied Measurement in Education, 6,* 269–279.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33,* 453–464.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Coffman, W. E. (1961). Sex differences in regard to items in an achievement test. In *Eighteenth yearbook: National Council on Measurement in Education* (pp. 117–124). Washington, DC: National Council on Measurement in Education.

Cohen, A. S., & Bolt, D. M. (2002). *A mixture model analysis of differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 237–255.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–220). New York: American Council on Education & Macmillan.

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

College Board. (1914). *Annual report.* New York: College Board.

Conant, James B. (1961). *Slums and Suburbs: A Commentary on Schools in Metropolitan Areas.* New York: McGraw Hill.

Coulehan, M. (2004). ACT now and then. *iJournal*, 8. Retrieved June 24, 2004, from http://www.ijournal.us/issue_08/ij_issue08_MichaelCoulehan_01.htm

Darlington, R. B. (1971). Another look at "culture fairness." *Journal of Educational Measurement, 8,* 71–82.

de Graaf, J. W. (1999). *Relating new to old*: *a classic controversy in developmental psychology.* University of Groningen (RUG, Ontwikkelingspsychologie en Experimentele Klinische Psychologie, BCN): Regenboog Drukkerij. (Doctoal dissertation thesis, Netherlands).

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–68.

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

Educational Testing Service. (2003). *Fairness review guidelines*. Princeton, NJ: Author.

Eells, K., Davis, A., Havighurst, R., Herrick, V., & Tyler, R. (1951). *Intelligence and cultural differences.* Chicago: University of Chicago Press.

Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin, 75,* 261–269.

Englehard, G. (1989). Accuracy of bias review judges in identifying teacher certification tests. *Applied Measurement in Education, 3,* 347–360.

Evinger, S. (1995). How shall we measure our nation's diversity? *Chance, 8*, 7–14.

Gardener, H. (1999). Who owns intelligence? *Atlantic Monthly, 283*, 67–76.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement, 63*, 65–74.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying and content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement, 40,* 281–306.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20,* 26–36.

Gipps, C. (1999). Socio-cultural aspects of assessment. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 355–392). Washington, DC: AERA.

Gould, Stephen Jay (1996). *The mismeasure of man: Revised and expanded.* New York: W. W. Norton. (Original work published 1981)

Green, D. R., & Draper, J. F. (1972). *Exploratory studies of bias in achievement tests.* Paper presented to the American Psychological Association, Honolulu, HI.

Griggs v. Duke Power Company, 401 U.S. 424 (1971).

Grutter v. Bollinger (02–241) 539 U.S. 306 (2003).

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications.* Hingham, MA: Kluwer, Nijhoff.

Hartigan, J. A. & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery*. Washington, DC: National Academy Press.

Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement, 62*, 32–44.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

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

Hu, P. G., & Dorans, N. L. (1989, March). *The effect of deleting items with extreme differential item functioning on equating functions and reported score distributions.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hubin, D. R. (1988). *The Scholastic Aptitude Test: Its development and introduction, 1900–1948.* Ph.D. dissertation, University of Oregon at Eugene.

Hunter, J. F. (1975, December). A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. A paper presented at the National Institute of Education Conference on Test Bias. Annapolis, MD.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16,* 209–225.

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 40,* 185–244.

Jensen, A. R. (1975). *Test bias and construct validity.* Invited address at the American Psychological Association, Chicago, September.

Jensen, A. R. (1980). *Test bias*. New York: Free Press.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education*. Washington, DC: Author.

Kim, S., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's areas measures, and the likelihood ration test on the detection of differential item functioning. *Applied Measurement in Education, 8*, 291–312.

Lan, W., Bradley, L., Tallent-Runnels, M., & Hsu, P.-Y. (2001, April). *Changes in student academic performance and perceptions of school and self before dropping out from schools.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Lawrence, I. M., Rigol, G. W., Van Essen, T., & Jackson, C. A. (2003). *A historical perspective on the content of the SAT.* New York: College Entrance Examination Board.

Lazerson, M. (2001). The College Board and American educational history. In M. C. Johanek (Ed.), *A faithful mirror: Reflections on the College Board and education in America* (pp. 379–400). New York: College Board.

Le, V.-N. (1999). *Identifying differential item functioning on the NELS:88 history achievement test.* Center for the Study of Evaluation, Los Angeles, CA: CRESST/UCLA.

Lemann, N. (2004). A history of admissions testing, In R. Zwick (Ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions* (pp. 5–14). New York: RoutledgeFalmer.

Lemann, N. (1995). The structure of success in America. *Atlantic Monthly, 276,* 41–60.

Lemann, N. (1999). *The big test*. New York: Farrar, Straus & Giroux.

Levin, S. (2003). Social psychological evidence on race and racism. In M. Chang, D. Witt, K. Haikuta, & J. Jones (Eds.), *Compelling interest: Examining the evidence on racial dynamics in higher education in colleges and universities* (pp. 97–125). Stanford University Press.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 119–158). Washington, DC: American Council on Education.

Lindquist, E. F. (1970). Iowa Testing Programs—A retrospective view. *Education, 91*, 7–23.

Linn, R. L. (1989). Current directions and future perspectives. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 1–10). New York: American Council on Education & Macmillan.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139–161.

Linn, R. L. (1982). Ability testing: Individual differences, prediction and differential prediction. In A. K. Wigdor, & W. R. Garner (Eds.), *Ability Testing: Uses, consequences and controversies, Part II* (pp. 335–388). Washington, DC: National Academy Press.

Lombardo, P. A. (2002). "The American Breed": Nazi Eugenics and the origins of the Pioneer Fund. *Albany Law Review, 65*(3), 743–830.

Lombardo, P. A. (2003). Facing Carrie Buck. *Hastings Center Report, 33*, 14–16.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH DIF across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam: Swets & Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690–700.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–748.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*, 15.

Masters, G. N. (1993). Undesirable item discrimination. Rasch Measurement Transactions, 7, 289.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*, 131–144.

Messick, S. (1994). Foundations of validity: Meaning and consequence in psychological assessment. *European Journal of Psychological Assessment 10*, 1–9.

Mill, J. S. (1848). *Principles of political economy.* (Reprinted in *Collected works of John Stuart Mill* [Vol. 2], Toronto: University of Toronto Press, 1965)

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scores items. *Journal of Educational Measurement, 30,* 107–122.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical procedures for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Moss, S. M. (2003). Conceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*(4), 13–25.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple group partial credit model. *Journal of Educational Measurement, 36*, 217–232.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30,* 293–311.

NCME Ad Hoc Committee on the Development of a Code of Ethics. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: National Council on Measurement in Education.

Novick, M. R., & Petersen, N. S. (1976). Toward equalizing educational and employment opportunity. *Journal of Educational Measurement, 13*, 77–88.

Office for Minority Education. (1980). *An approach for identifying and minimizing bias in standardized tests* (Educational Testing Service Monograph No. 4). Princeton, NJ: ETS.

Office of Technology Assessment, U.S. Congress. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

OMB. (1997, October 30). Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, Federal Register Notice (62FR58782-89). Washington, DC: Author.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In

P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Penfield, R. D. (2001). Assessing differential item functioning across multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*, 235–259.

Penfield, R. D., & Camilli, G. (in press). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics*: *Psychometrics*, 25. North Holland: Elsevier.

Penfield, R. D., & Lam, T. C. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5–15.

Penny, J., & Johnson, R. L. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel chi-square DIF index. *Journal of Experimental Education*, *67*, 343–366.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture fair selection. *Journal of Educational Measurement, 13,* 3–29.

Peterson, J. J. (1983). *The Iowa Testing Programs: The first fifty years*. Iowa City, IA: University of Iowa Press.

Phillips, A. & Holland, P.W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics*, 425-431.

Popham, W. J. (2001). Standardized achievement tests: Misnamed and misleading. *Education Week, 21*(03), 46. (www.edweek.org/ew/newstory.cfm?slug=03popham.h21)

Porter, T. M. (1986). *The rise of statistical thinking 1820–1900.* Princeton, NJ: Princeton University Press.

Prowker, A., & Camilli, G. (2004. *Beyond the composite: An item level methodological study of NAEP mathematics results.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 492–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report No. 93-1). New York: College Board.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Hillsdale, NJ: Erlbaum.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29,* 1–41.

Rawls, J. (1971). *A theory of justice.* Cambridge, MA: Harvard University Press.

Raymond, M. R. & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives* (pp. 119–158). Mahwah, N.J.: Erlbaum.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25,* 193–203.

Reese, L., Balzano, S., Gallimore, R., & Goldenberg, C. (1995). The concept of educación: Latino family values and American schooling. *International Journal of Educational Research, 23*, 57–81.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178–201). New York: Wiley.

Roe v. Wade, 410 U.S. 113 (1973).

Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Roussos, L. A., & Stout, W. F. (2004). Differential item functioning analysis: Detecting DIF items and testing DIF hypotheses. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 107–115)**.** Thousand Oaks, CA: Sage.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*, 248–270.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143–152.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale, NJ: Lawrence Erlbaum.

Shealy, R., & Stout, W. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Shealy, R., & Stout, W. (1993b). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–240). Hillsdale, NJ: Erlbaum.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 1–14.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 1066–1101) (4th ed.). Washington, DC: American Educational Research Association.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external criteria. *Journal of Educational Statistics*, *6,* 317–375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93–128.

Smedley, A. (2002). Science and the idea of *race*. In J. M. Fish (Ed.), *Race and intelligence* (pp. 145–176). Mahwah, NJ: Erlbaum.

Spray, J. A. (1989). *Performance of three conditional DIF statistics in detecting differential item functioning on simulated tests* (ACT Research Report Series 89-7). Iowa City, IA: ACT.

Stiggins, R. J. (2002). Where is our assessment future and how can we get there from here? In R. W. Lissitz & W. D. Schafer (Eds.), *Assessment in educational reform* (pp. 18–48). Boston: Allyn and Bacon.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Takala, S., & Kaftandieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*(3), 323–340.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Detection of differential item functioning using the parameters of item response theory models. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Use of item response theory in the study of group differences in trace lines. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 147–169). Hillsdale, NJ: Erlbaum.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, *8,* 63–70.

Thurstone, L. L. (1925). A method of scaling educational and psychological tests. *Journal of Educational Psychology, 8,* 63–70.

Thurstone, L. L. (1931). Influence of motion pictures on children's attitudes. *Journal of Social Psychology, 2,* 291–305.

Walker, C. M., Beretvas, S. N., & Ackerman, T. A. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. *Applied Measurement in Education, 14*, 3–16.

Wechsler, H. S. (2001). Eastern standard time: High-school college collaboration and admission to college 1880–1930. In M. C. Johanek (Ed.), *A faithful mirror: Reflections on the College Board and education in America* (pp. 41–79). New York: College Board.

Welner, K. (2001). Alexander v. Sandoval: A setback for civil rights. *Educational Policy Analysis Archives, 24.* Retrieved September 12, 2004, from http://epaa.asu.edu/epaa/v9n24.html

Westers, P., & Kelderman, H. (1991). Examining differential item functioning due to item difficulty and alternative attractiveness. *Psychometrika, 57*, 107–118.

Wigdor, A. K. & Sackett, P. R. (1993). Employment testing and public policy: The case of the General Aptitude Test Battery. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 183–204). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wightman, L. F. (2000). The role of standardized tests in the debate about merit, academic standards, and affirmative action. *Psychology, Public Policy, and Law, 6*, 90–100.

Wightman, L. (2003). Standardized testing and equal access: A tutorial. In M. Chang, D. Witt, K. Haikuta, & J. Jones (Eds.), *Compelling interest: Examining the evidence on racial dynamics in higher education in colleges and universities* (chap. 4). Stanford, CA: Stanford University Press.

Wright, B. D., Mead, R., & Drada, D. (1976). *Detecting and correcting item bias with a logistic response model* (Mesa Research Memorandum 22). Chicago: University of Chicago, MESA Psychometric Laboratory.

Young, J. W. (2001). *Differential validity, differential prediction, and college admissions testing: A comprehensive review and analysis* (Research Report No. 2001-6). New York: The College Board.

Young, M. D. (1958). *Rise of the meritocracy.* Baltimore: Penguin Books.

Young, M.D. (2001, June 29). Down with Meritoocracy. *Guardian Unlimited*. Retrieved August 1, 2005 from http://www.guardian.co.uk/comment/story/0,3604,514207,00.html.

Zenderland, L. (1998). *Measuring minds: Henry Herbert Goddard and the origins of American intelligence testing.* Cambridge: Cambridge University Press.

Zenisky, A., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large scale assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*, 541–564.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores* [On-line]. Ottawa, Ontario, Canada: Department of National Defense, Directorate of Human Resources Research and Evaluation. Available: http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–198.

Zwick, R. (2002). *Fair game?* New York: Routledge Falmer.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Zwick, R. Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36, 1–28.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10,* 321–334.

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32* (4), 341–363.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121–140.

611-190-007.indd 256 7/14/2006 4:18:28 PM