The Problem of Bias in Psychological Assessment

15

Test bias: In God we trust; all others must have data.

Reynolds (1983)

Abstract

Much the impetus for the current debate about bias in psychological testing is based on well-documented, consistent, and substantive differences between IQ scores of Whites, Hispanics, and Blacks in the U.S.A. Various explanations are offered for these differences including the idea that IQ tests are inherently biased against Blacks, Hispanics, and possibly other ethnics groups, or what is commonly known as the Cultural Test Bias Hypothesis (CTBH). Because tests are used to make many different and important decisions about people, lack of fairness in testing resulting from test bias is of grave concern. This chapter traces the historical roots of the CTBH to the present day, provides important distinctions regarding different definitions of test bias that are critical for empirical examination of the issue, presents common objections to the use of psychological testing, and describes how test authors and publishers detect bias in psychological tests. The chapter concludes by noting that while more research is necessary, the current evidence largely supports the proposition that most commercially developed widely use tests of achievement and aptitude are not culturally biased.

Supplementary Information The online version of this chapter (https://doi.org/10.1007/978-3-030-59455-8_15) contains supplementary material, which is available to authorized users.

C. R. Reynolds et al., *Mastering Modern Psychological Testing*,
https://doi.org/10.1007/978-3-030-59455-8_15

Learning Objectives
After reading and studying this chapter, students should be able to:

1. Explain the cultural test bias hypothesis.
2. Describe alternative explanations for observed group differences in performance on aptitude and other standardized tests.
3. Describe the relationship between bias and reliability.
4. Describe the major objections regarding the use of standardized tests with minority students.
5. Describe what is meant by cultural loading, cultural bias, and culture--free tests.
6. Describe the mean difference definition of test bias and its current status.
7. Describe the results of research on the presence of bias in the content of educational and psychological tests.
8. Describe the results of research on the presence of bias in other internal features of educational and psychological tests.
9. Describe the results of research on bias in prediction and in relation to variables that are external to the test.
10. Explain what is implied by homogeneity of regression and describe the conditions that may result when it is not present.

Much effort has been expended to determine why group differences occur on standardized aptitude tests, but we do not know for certain why.

Groups of people who can be defined based on descriptors such as gender or ethnicity do not always perform that same way on educational and psychological tests. For example, on tests of spatial skill, requiring visualization and imagery, men and boys tend to score higher than do women and girls. On tests that involve written language and tests of simple psychomotor speed, women and girls tend to score higher than men and boys (see Special Interest Topic 15.1 for additional information). Ethnic group differences in test performance also occur and are the most controversial and polemic of all group differences.

The cultural test bias hypothesis holds that differences in mean test scores across gender or ethnic groups are due to artifacts of the test or measurement process and do not reflect real differences among groups on the constructs or dimensions purported to be measured.

There is perhaps no more controversial finding in the field of psychology than the persistent one standard deviation difference between the intelligence test performance of Black and White students taken as a group, which is 15 standard score points on most IQ tests. There are many, many such group differences on various measures of specialized ability and achievement—and these differences go in

Special Interest Topic 15.1: Sex Differences in Intelligence

Research has shown that although there are no significant sex differences in overall intelligence scores, substantial differences exist with regard to specific cognitive abilities. Females typically score higher on a number of verbal abilities whereas males perform better on visual-spatial and (starting in middle childhood) mathematical skills. It is believed that sex hormone levels and social factors both influence the development of these differences. As is typical of group differences in intellectual abilities, the variability in performance within groups (i.e., males and females) is much larger than the mean difference between groups (Neisser et al., 1996). Diane Halpern (1997) has written extensively on gender differences in cognitive abilities. This table briefly summarizes some of her findings.

Selected abilities on which women obtain higher average scores

Type of ability	Examples
Rapid access and use of verbal and other information in long-term memory	Verbal fluency, synonym generation, associative memory, spelling, anagrams
Specific knowledge areas	Literature and foreign languages
Production and comprehension of prose	Writing and reading comprehension
Fine motor tasks	Matching and coding tasks, pegboard, mirror tracing
School performance	Most subjects

Selected abilities on which men obtain higher average scores

Type of ability	Examples
Transformations of visual working memory, moving objects, and aiming	Mental rotations, dynamic spatiotemporal tasks, accuracy in throwing
Specific knowledge areas	General knowledge, mathematics, science, and geography
Fluid reasoning	Proportional, mechanical, and scientific reasoning; SAT Math and GRE Quantitative

Source: This table was adapted from Halpern (1997, Appendix, p. 1102)

both directions. Much effort has been expended to determine why group differences occur but we do not know for certain why they exist. One major, carefully studied explanation is that the tests are biased in some way against certain groups. This is referred to as the ***CTBH*** (Cultural Test Bias Hypothesis).

The CTBH represents the contention that any gender, ethnic, racial, or other nominally determined groups who perform differently on mental tests are due to inherent, artifactual biases produced within the tests through flawed psychometric methodology. Group differences are believed then to stem from characteristics of the tests and to be unrelated to any actual differences in the psychological trait, skill,

or ability in question. The resolution or evaluation of the validity of the CTBH is one of the most crucial scientific questions facing psychology today.

Bias in mental tests has many implications for individuals including the misplacement of students in educational programs, errors in assigning grades, unfair denial of admission to college, graduate, and professional degree programs, and the inappropriate denial of employment. The scientific implications are even more substantive. There would be dramatic implications for educational and psychological research and theory if the CTBH were correct: The principal research of the past 100 years in the study of the psychology of human differences would have to be dismissed as confounded and the results deemed largely artifactual because much of the work is based on standard psychometric theory and testing technology. This would in turn create major upheavals in professional psychology, because the foundations of clinical, counseling, educational, industrial, and school psychology are all strongly tied to the basic academic field of individual differences.

Each day psychologists in clinical practice use psychological tests to make diagnostic decisions that affect the lives of their patients in many ways, e.g., treatment approaches, types of psychopharmacological agents that may be applied, suitability for employment, and in forensic settings, even eligibility to receive the death penalty is affected by the results of intelligence tests. School psychologists arrive at diagnostic and eligibility decisions that determine school placements. Industrial and organizational psychologists design screening programs that test job applicants for employment skills and screen public safety officer applicants for various personality traits that predict success in law enforcement. Educational psychologists conduct research that assesses outcomes in learning environments using standardized tests in order to determine what methods and environments for learning are the most successful. These are examples of but a few of the many uses of psychological tests in the everyday practice of psychology. Typically, professionally designed tests used for such decision-making are subjected to lengthy development stages and tryout periods and are held up to stringent psychometric and statistical standards. If these methods turn out to be culturally biased when used with native-born American ethnic minorities, what about other alternative methods of making these decisions that are inherently more subjective, e.g., interviews, observation, review of references, performance, or portfolio assessments? Put another way, if well-constructed and properly standardized tests are biased, then less standardized, more subjective approaches are almost certain to be at least as biased and probably more so. As the reliability of a test or evaluation procedure goes down, the likelihood of bias goes up, the two being inversely related. A large reliability coefficient does not eliminate the possibility of bias, but as reliability is lowered, the probability that bias will be present increases.

If well-constructed and properly standardized tests are biased, then interviews and other subjective evaluation procedures are almost certain to be at least as biased and probably more so.

The purpose of this chapter is to address the issues and findings surrounding the CTBH in a rational manner and evaluate the validity of the hypothesis, as far as

possible, on the basis of existing empirical research. This will not be an easy task because of the controversial nature of the topic and its strong emotional overtones. Prior to turning to the reasons that test bias generates highly charged emotions and reviewing some of the history of these issues, it is proper to engage in a discussion of just what we mean by the term bias.

15.1 What Do We Mean by Bias?

Bias carries many different connotations for the lay public and for professionals in a number of disciplines.

Bias carries many different connotations for the lay public and for professionals in a number of disciplines. To the legal mind, bias denotes illegal discriminatory practices while to the lay mind it may conjure notions of prejudicial attitudes. Much of the rancor in psychology and education regarding proper definitions of ***test bias*** is due to the divergent uses of this term in general but especially by professionals in the same and related academic fields.

As presented in the *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), bias is discussed within the overall context of fairness, a multi-faceted and complex concept that can include both subjective and objective arguments across a variety of both professional and casual discussions. Up to this point, we have directly or indirectly infused the notion of fairness throughout this book, whether it be in the context of:

- *Interpreting test scores*, e.g., use norm or criterion-referenced interpretations
- *Minimizing error in test scores*, e.g., follow standardized procedures for all examinees
- *Reducing threats to validity of score interpretations*, e.g., provide appropriate accommodations to those with disabilities
- *Writing appropriate test items*, e.g., conduct expert review to ensure cultural sensitivity and design to measure intended construct across a wide range of test-taker ability
- *Evaluating potential of job candidates*, e.g., use same standard criteria for all applicants
- *Disability evaluation*, e.g., establish diagnosis of intellectual disability using best assessment procedures

Fairness is fundamental to the development and use of psychological tests and the subsequent interpretation of scores. Fairness is required during the testing process so that all examinees are given a similar opportunity to demonstrate their true ability on the construct being assessed. Factors irrelevant to the construct are eliminated during assessment ensuring the construct is measured in a way that is impacted only by knowledge, skills, or abilities relevant to the construct itself. Standardized test materials, instructions, administration, and scoring also promote fairness, although there are some exceptions. For example,

consider a parent of child being evaluated in school for a learning disability. The parent is an English-language learner who has a reasonable mastery of the language, but is unable to read effectively. As part of the child's evaluation, the parent is asked to complete a rating form designed to assess the behavioral-emotional functioning of the child. To ensure only the construct of interest is measured (i.e., child's functioning), the parent is allowed to complete the rating form by responding to an audio recording of the questions because being able to read the test item is not a requirement of assessing the behavioral-emotional functioning of the child.

Fairness can also be characterized by the extent to which the characteristics of the test itself are related to the construct being measured, and not to characteristics that are associated with attributes of a specific group. In the statistical sense, this is bias. The *Standards* defines bias as—a systematic error in a test score (AERA et al., 2014). A biased assessment is one that systematically underestimates or overestimates the value of the variable it is designed to measure. If the bias is a function of a nominal cultural variable (e.g., ethnicity or gender), then the test has a cultural bias. As an example, if an achievement test produces different mean scores for different ethnic groups, and there actually are true differences between the groups in terms of achievement, the test is not biased. However, if the observed differences in achievement scores are the result of the test underestimating the achievement of one group or overestimating the achievement of another, then the test is culturally biased.

> In terms of tests and measurements, bias is something systematic that distorts construct measurement or prediction by test scores of other important criteria.

Other definitions of the term bias in research on the CTBH or cross-group validity of tests are unacceptable from a scientific perspective for two reasons: (1) The imprecise nature of other uses of the term bias makes empirical investigation and rational inquiry exceedingly difficult, and (2) other uses of the term invoke specific moral value systems that are the subject of intense emotional debates that do not have a mechanism for rational resolution. It is imperative that the evaluation of bias in testing be undertaken from the standpoint of scholarly inquiry and debate. Emotional appeals, legal-adversarial approaches, and political remedies of scientific issues appear to us to be inherently unacceptable and unuseful.

15.2 Past and Present Concerns: A Brief Look

Concern about cultural bias in mental testing has been a recurring issue since the beginning of the use of assessment in education. From Alfred Binet in the 1800s to Arthur Jensen over the last 50 years, many scientists have addressed this controversial problem, with varying, inconsistent outcomes. In the last few decades, the issue of cultural bias has come forth as a major contemporary problem far exceeding the bounds of purely academic debate and professional rhetoric. The debate over the CTBH has become entangled and sometimes confused within the larger issues of

individual liberties, civil rights, and social justice, becoming a focal point for psychologists, sociologists, educators, politicians, minority activists, and the lay public. The issues increasingly have become legal and political. Numerous court cases have been brought and New York state even passed "truth-in-testing" legislation that is being considered in other states and in the federal legislature. Such attempts at solutions are difficult. Take for example the legal response to the question "Are intelligence tests used to diagnose intellectual disability biased against cultural and ethnic minorities?" In California in 1979 (Larry P. v. Riles) the answer was "yes" but in Illinois, in 1980 (PASE v. Hannon), the response was "no." Thus two federal district courts of equivalent standing have heard nearly identical cases, with many of the same witnesses espousing much the same testimony, and reached precisely opposite conclusions. See Special Interest Topic 15.2 for more information on legal issues surrounding assessment bias.

Though current opinion on the CTBH is quite divergent, ranging from those who consider it to be for the most part unresearchable (e.g., Schoenfeld, 1974) to those who considered the issue settled decades ago (e.g., Jensen, 1980), it seems clear that empirical analysis of the hypothesis should continue to be undertaken. However difficult full objectivity may be in science, we must make every attempt to view all socially, politically, and emotionally charged issues from the perspective of rational scientific inquiry. We must also be prepared to accept scientifically valid findings as real, whether we like them or not.

15.3 The Controversy Over Bias in Testing: Its Origin, What It Is, and What It Is Not

Systematic group differences on standardized intelligence and aptitude tests may occur as a function of socioeconomic level, race or ethnic background, and other demographic variables. Black-White differences on IQ measures have received extensive investigation for more than 50 years. Although results occasionally differ slightly depending on the age groups under consideration, random samples of Blacks and Whites show a mean difference of about 1 standard deviation, with the mean score of the White groups consistently exceeding that of the Black groups. When a number of demographic variables are taken into account (most notably socioeconomic status, or SES), the size of the difference reduces to 0.5–0.7 standard deviation but remains robust. The differences have persisted at relatively constant levels for quite some time and under a variety of methods of investigation. Some research suggests these gaps are narrowing (Dickens & Flynn, 2006; Neisser et al., 1996; Nisbett, 2009), while other research disputes the narrowing of gaps (Rushton & Jensen, 2005, 2010).

Mean differences between ethnic groups are not limited to Black-White comparisons. Although not nearly as thoroughly researched as Black-White differences, Hispanic-White differences have also been documented, with Hispanic mean performance approximately 0.5 standard deviation below the mean of the White group. On average, Native Americans tend to perform lower on tests of verbal intelligence

Special Interest Topic 15.2: Courtroom Controversy Over IQ Testing in the Public Schools

Largely due to overall mean differences in the performance of various ethnic groups on IQ tests, the use of intelligence tests in the public schools has been the subject of courtroom battles around the United States. Typically such lawsuits argue that the use of intelligence tests as part of the determination of eligibility for special education programs leads to overidentification of certain minorities (traditionally African American and Hispanic children). A necessary corollary to this argument is that the resultant overidentification is inappropriate because the intelligence tests in use are biased, underestimating the intelligence of minority students, and that there is in fact no greater need for special education placement among these ethnic minorities than for other ethnic groups in the population.

Attempts to resolve the controversy over IQ testing in the public schools via the courtroom have not been particularly successful. Unfortunately, but not uncharacteristically, the answer to the legal question "Are IQ tests biased in a manner that results in unlawful discrimination against minorities when used as part of the process of determining eligibility for special education placements?" depends on where you live! There are four key court cases to consider when reviewing this question, two from California and one each from Illinois and Georgia.

The first case is Diana v. State Board of Education (C-70-37 RFP, N.D. Cal., 1970), heard by the same federal judge who would later hear the Larry P. case (see later discussion). Diana was filed on behalf of Hispanic (referred to as Chicano at that time and in court documents) children classified as EMR, or educable mentally retarded (a now archaic term that has been replaced with intellectual disability), based on IQ tests administered in English. However, the children involved in the suit were not native English speakers and when retested in their native language, all but one (of nine) scored above the range designated as EMR. Diana was resolved through multiple consent decrees (agreements by the adverse parties ordered into effect by the federal judge). Although quite detailed, the central component of interest here is that the various decrees ensured that children would be tested in their native language, that more than one measure would be used, and that adaptive behavior in nonschool settings would be assessed prior to a diagnosis of EMR.

It seems obvious to us now that whenever persons are assessed in other than their native language, the validity of the results as traditionally interpreted would not hold up, at least in the case of ability testing. This had been obvious to the measurement community for quite some time prior to Diana, but it had not found its way into practice. Occasionally one still encounters cases of a clinician evaluating children in other than their native language and making inferences about intellectual development—clearly this is inappropriate.

Three cases involving intelligence testing of Black children related to special education placement went to trial: Larry P. v. Riles (343 F. Supp. 306,

1972; 495 F. Supp. 976, 1979); PASE v. Hannon (506 F. Supp. 831, 1980); and Marshall v. Georgia (CV 482-233, S.D. of Georgia, 1984). Each of these cases involved allegations of bias in IQ tests that caused the underestimation of the intelligence of Black children and subsequently led to disproportionate placement of Black children in special education programs. All three cases presented testimony by experts in education, testing, measurement, and related fields, some professing the tests to be biased and others professing they were not. That a disproportionate number of Black children were in special education was conceded in all cases—what was litigated was the reason.

In California in Larry P. v. Riles (Wilson Riles being superintendent of the San Francisco Unified School District), Judge Peckham ruled that IQ tests were in fact biased against Black children and resulted in discriminatory placement in special education. A reading of Peckham's decision reveals a clear condemnation of special education, which is critical to Peckham's logic. He determined that because special education placement was harmful, not helpful, to children, the use of a test (i.e., IQ) that resulted in disproportionate placement was therefore discriminatory. He prohibited (or enjoined) the use of IQ tests with Black children in the California public schools.

In PASE v. Hannon (PASE being an abbreviation for Parents in Action on Special Education), a similar case to Larry P. was brought against the Chicago public schools. Many of the same witnesses testified about many of the same issues. At the conclusion of the case, Judge Grady ruled in favor of the Chicago public schools, finding that although a few IQ test items might be biased, the degree of bias in the items was inconsequential.

In Marshall v. Georgia, the NAACP brought suit against rural Georgia school districts alleging bias in the instructional grouping and special education placement associated with IQ testing. Although some of the same individuals testified in this case, several new opinions were offered. However, the judge in Marshall eventually ruled in favor of the schools, finding that IQ tests were not in fact biased, and that a greater actual need for special education existed in minority populations.

In the courtroom, we are no closer to resolution of these issues today than we were in 1984 when Marshall was decided. However, these cases and other societal factors did foster much research that has brought us closer to a scientific resolution of the issues. They also prompted the development of new, up-to-date IQ tests and more frequent revisions or updating of older tests. Many challenges remain, especially that of understanding the continued higher failure rates (relative to the majority ethnic population of the United States) of some ethnic minorities in the public schools (while other ethnic minorities have a success rate that exceeds the majority population) and the disproportionate referral rates by teachers of these children for special education placement. The IQ test seems to be only one of many messengers in this crucial educational issue, and bias in the tests does not appear to be the answer.

than Whites. Both Hispanics and Native Americans tend to perform better on visual--spatial tasks relative to verbal tasks. All studies of race/ethnic group differences on ability tests do not show higher levels of performance by Whites. Asian American groups have been shown consistently to perform as well as or better than White groups. Depending on the specific aspect of intelligence under investigation, other race/ethnic groups show performance at or above the performance level of White groups (for a readable review of this research, see Neisser et al., 1996).

It should always be kept in mind that the overlap among the distributions of intelligence test scores for different ethnic groups is much greater than the size of the differences between the various groups. Put another way, there is always more within-group variability in performance on mental tests than between-group variability. Neisser et al. (1996) frame it this way:

> Group means have no direct implications for individuals. What matters for the next person you meet (to the extent that test scores matter at all) is that person's own particular score, not the mean of some reference group to which he or she happens to belong. The commitment to evaluate people on their own individual merit is central to a democratic society. It also makes quantitative sense. The distributions of different groups inevitably overlap, with the range of scores within any one group always wider than the mean differences between any two groups. In the case of intelligence test scores, the variance attributable to individual differences far exceeds the variance related to group membership. (p. 90)

15.3.1 Explaining Mean Group Differences

Once mean group differences are identified, it is natural to attempt to explain them. Reynolds (2000) notes that the most common explanations for these differences have typically fallen into four categories:

1. The differences primarily have a genetic basis
2. The differences have an environmental basis (e.g., SES, education, culture)
3. The differences are due to the interactive effect of genes and environment
4. The tests are defective and systematically underestimate the knowledge and skills of minorities

The final explanation (i.e., Category 4) is embodied in the CTBH introduced earlier in this chapter. Restated, the CTBH represents the contention that any gender, ethnic, racial, or other nominally determined group differences on mental tests are due to inherent, artifactual biases produced within the tests through flawed psychometric methodology. Group differences are believed then to stem from characteristics of the tests and to be totally unrelated to any actual differences in the psychological trait, skill, or ability in question. Because mental tests are based largely on middle-class values and knowledge, their results are more valid for those groups and will be biased against other groups to the extent that they deviate from those values and knowledge bases. Thus, ethnic and other group differences result from flawed psychometric methodology and not from actual differences in aptitude. As will be discussed, this hypothesis reduces to one of ***differential validity***; the

hypothesis of differential validity being that tests measure intelligence and other constructs more accurately and make more valid predictions for individuals from the groups on which the tests are mainly based than for those from other groups. The practical implications of such bias have been pointed out previously and are the issues over which most of the court cases have been fought.

If the CTBH is incorrect, then group differences are not attributable to the tests and must be due to one of the other factors mentioned previously. The model emphasizing the interactive effect of genes and environment (category c, commonly referred to as the Environment × Genetic Interaction Model) is dominant among contemporary professionals who reject the argument that group differences are artifacts of test bias; however, there is much debate over the relative contributions of genetic and environmental factors (Reynolds, 2000; Suzuki & Valencia, 1997). In addition to the models noted, Williams (1970), Helms (1992), and Richardson (1993) proposed other models with regard to Black-White differences on aptitude tests, raising the possibility of qualitatively different cognitive structures that require different methods of measurement.

The controversy over test bias should not be confused with that over etiology of any observed group differences.

15.3.2 Test Bias and Etiology

The controversy over test bias is distinct from the question of etiology. Reynolds and Ramsay (2003) note that the need to research etiology is only relevant once it has been determined that mean score differences are real, not simply artifacts of the assessment process. Unfortunately, measured differences themselves have often been inferred to indicate genetic differences and therefore the genetically based intellectual inferiority of some groups. This inference is not defensible from a scientific perspective.

15.3.3 Test Bias and Fairness

As mentioned previously, bias is considered a portion of the overall construct of fairness. As noted by Brown, Reynolds, and Whitaker (1999), fairness is a moral, philosophical, or legal issue on which reasonable people can disagree. On the other hand, bias is a statistical property of a test. Therefore, bias is a property empirically estimated from test data whereas fairness is a principle established through debate and opinion. Nevertheless, it is common to incorporate information about bias when considering the fairness of an assessment process. For example, a biased test would likely be considered unfair by essentially everyone. However, it is clearly possible that an unbiased test might be considered unfair by at least some. Special Interest Topic 15.3 summarizes the discussion of fairness in testing and test use from the *Standards* (AERA et al., 2014).

Special Interest Topic 15.3: Fairness and Bias: A Complex Relationship

The Standards (AERA et al., 2014) present four different ways that fairness is typically used in the context of assessment.

1. *Fairness in treatment during the testing process*: A primary goal of testing is to maximize the opportunity for test takers to demonstrate their knowledge or ability on the construct being measured. Carefully developed tests that follow standardized administration procedures in a controlled environment suitable for completing the test achieve this goal. Factors such as proper seating, adequate lighting, strictly controlled time limits, and test proctor responsibilities can be adequately controlled with minimal effort within a test setting, but for nationally based tests that occur (e.g., the SAT, the National Council Licensure Examination for nurses, etc.), it can be harder to control across multiple settings. Differences across settings can lead to inadvertent advantages for some test takers over others. Thus, to enable fairness across such settings, it is important to establish guidelines for use across multiple settings.
2. *Fairness as a lack of measurement bias*: When a test measures an attribute unrelated to the intended construct being measured or the manner in which the test is used, it can result test score differences across subgroups. Differential item functioning (DIF) occurs when test takers of equal ability do not have the same probability of answering a test item correctly. An indication of DIF must be accompanied by a suitable, substantial explanation for DIF to justify an item is biased. Differential test functioning (DTF) refers to differences in the functioning of tests between defined groups. DTF indicates that individuals from different groups who have the same standing on the construct being measured do not have the same expected test score. Items that indicate DIF or test scores that indicate DTF can lead to predictive bias, which is found when differences exist in the pattern of associations between test scores and other variables for different groups, causing concerns about bias in the inferences drawn from the use of test scores. Regression is used to determine differential prediction is present, which can be measured by slope and/or intercept differences in the regression analyses between targeted groups.
3. *Fairness in access to the construct as measured*: A goal of fairness in testing is to allow all test takers an opportunity to demonstrate their standing on the construct being measured. Accessible testing situations enable test takers to show their status on the construct without being unduly advantaged or disadvantaged by irrelevant individual characteristics (e.g., age, race, disability status, gender, etc.). Accessibility can be understood when contrasting the knowledge, skills, and abilities that reflect the construct being measured by the test with the knowledge, skills, and abilities that are required for test takers to respond to the test tasks or items. Factors related

to individual characteristics can restrict accessibility can interfere with measuring the construct. For example, presenting a personality test in Braille or large-print format makes it more accessible to a visually impaired person, increasing the changes for a valid measurement of a person's personality characteristics. When test-taker characteristics that impede accessibility are related to the construct being measured (e.g., dyslexia and a test of reading), then adaptation of the construct might be warranted, and might result in more accurate measurement of the construct for the test taker, even if it is not directly comparable to the measurement of the original construct.

4. *Fairness as validity of individual test score interpretations for the intended uses*: Fairness is concerned with the validity of interpreting individual scores for their intended uses. While treating all individuals as similarly as possible is an important aspect of fairness, it is also important to take into account the individual characteristics of the test taker and understand how these characteristics may interfere with contextual factors of the testing situation and the interpretation of test scores. Test professionals are tasked with developing an understanding of when standardized testing procedures can and should be modified to obtaining more accurate measurement for certain groups of test takers, and when modifications can lead to an unfair advantage over other test takers.
5. In concluding the discussion of fairness, the *Standards* notes that fairness should not be perceived as equality of testing outcomes for relevant population subgroups. While differences in subgroups scores should increase scrutiny that possible test bias exists, group differences alone do not indicate that a testing application is biased or unfair.

15.3.4 Test Bias and Offensiveness

There is also a distinction between test bias and item offensiveness. Test developers often use a minority review panel to examine each item for content that may be offensive or demeaning to one or more groups (e.g., see Reynolds & Kamphaus, 2003, for a practical example). This is a good procedure for identifying and eliminating offensive items, but it does not ensure that the items are not biased. Research has consistently found little evidence that one can identify, by personal inspection, which items are biased and which are not (for reviews, see Camilli & Shepard, 1994; Reynolds, Lowe, & Saenz, 1999).

15.3.5 Test Bias and Inappropriate Test Administration and Use

The controversy over test bias is also not about blatantly inappropriate administration and usage of mental tests. Administration of a test in English to an individual for whom English is a poor second language is inexcusable both ethically and

legally, regardless of any bias in the tests themselves (unless of course, the purpose of the test is to assess English-language skills). It is of obvious importance that tests be administered by skilled and sensitive professionals who are aware of the factors that may artificially lower an individual's test scores. That should go without saying, but some court cases involve just such abuses. Considering the use of tests to assign pupils to special education classes or other programs, the question needs to be asked, "What would you use instead of a test?" Teacher recommendations alone are less reliable and valid than standardized test scores and are subject to many external influences. Whether special education programs are of adequate quality to meet the needs of children is an important educational question, but separate from the test bias question, a distinction sometimes confused.

15.3.6 Bias and Extraneous Factors

The controversy over the use of mental tests is complicated further by the fact that resolution of the cultural test bias question in either direction will not resolve the problem of the role of nonintellective factors that may influence the test scores of individuals from any group, minority, or majority. Regardless of any group differences, it is individuals who are tested and whose scores may or may not be accurate. Similarly, it is individuals who are assigned to classes and accepted or rejected for employment or college admission. Most assessment professionals acknowledge that a number of emotional and motivational factors may impact performance on intelligence tests. The extent to which these factors influence individual as opposed to group performance is difficult to determine.

15.4 Cultural Bias and the Nature of Psychological Testing

The question of cultural bias in testing arises from and is continuously fueled by the very nature of psychological and educational processes and how we measure those processes. Psychological processes are by definition internal to the organism and not subject to direct observation and measurement but must instead be inferred from behavior. It is difficult to determine one-to-one relationships between observable events in the environment, the behavior of an organism, and hypothesized underlying mediational processes. Many classic controversies over theories of learning revolved around constructs such as expectancy, habit, and inhibition. Disputes among different camps in learning were controversial and of long duration. Indeed, there are still disputes as to the nature and number of processes such as emotion and motivation. It should be expected that intelligence, as one of the most complex psychological processes, would involve definitional and measurement disputes that prove difficult to resolve.

In contrast, there are few charges of bias relating to physical measures that are on absolute scales, whether interval or ratio. Group differences in height, as an extreme example, are not attributed by anyone to any kind of cultural test bias. There is no

question concerning the validity of measures of height or weight of anyone in any culture. Nor is there any question about one's ability to make cross-cultural comparisons of these absolute measures.

The issue of cultural bias arises because of the procedures involved in psychological testing. Psychological tests measure traits that are not directly observable, subject to differences in definition, and measurable only on a relative scale. From this perspective, the question of cultural bias in mental testing is a subset, obviously of major importance, of the problem of uncertainty and possible bias in psychological testing generally. Bias might exist not only in mental tests but in other types of psychological tests as well, including personality, vocational, and psychopathological. Making the problem of bias in mental testing even more complex, not all mental tests are of the same quality; some are certainly psychometrically superior to others. There is a tendency for critics and defenders alike to overgeneralize across tests, lumping virtually all tests together under the heading mental tests or intelligence tests. Professional opinions of mental tests vary considerably, and some of the most widely used tests are not well respected by psychometricians. Thus, unfortunately, the question of bias must eventually be answered on a virtually test-by-test basis.

The question of bias must eventually be answered on a virtually test-by-test basis.

15.5 Objections to the Use of Educational and Psychological Tests with Minority Students

In 1969, the ***Association of Black Psychologists*** (ABPsi) adopted the following official policy on educational and psychological testing (Williams, Dotson, Dow, & Williams, 1980):

> The Association of Black Psychologists fully supports those parents who have chosen to defend their rights by refusing to allow their children and themselves to be subjected to achievement, intelligence, aptitude and performance tests which have been and are being used to (a) label Black people as uneducable; (b) place Black children in "special" classes and schools; (c) perpetuate inferior education in Blacks; (d) assign Black children to lower educational tracks than Whites; (e) deny Black students higher educational opportunities; and (f) destroy positive intellectual growth and development of Black people.

Since 1968 the ABPsi (a group with a current membership of about 1400) has sought a moratorium on the use of all psychological and educational tests with students from disadvantaged backgrounds. The ABPsi carried its call for a moratorium to other professional organizations in psychology and education. In direct response to the ABPsi call, the American Psychological Association's (APA) Board of Directors requested its Board of Scientific Affairs to appoint a group to study the use of psychological and educational tests with disadvantaged students. The committee report (Cleary, Humphreys, Kendrick, & Wesman, 1975) was subsequently published in the official journal of the APA, *American Psychologist*.

Subsequent to the ABPsi's policy statement, other groups adopted similarly stated policy statements on testing. These groups included the National Association for the Advancement of Colored People (NAACP), the National Education Association (NEA), the National Association of Elementary School Principals (NAESP), the American Personnel and Guidance Association (APGA), and others. The APGA called for the Association of Measurement and Evaluation in Guidance (AMEG), a sister organization, to "develop and disseminate a position paper stating the limitations of group intelligence tests particularly and generally of standardized psychological, educational, and employment testing for low socioeconomic and underprivileged and non-White individuals in educational, business, and industrial environments." It should be noted that the statements by these organizations assumed that psychological and educational tests are biased, and that what is needed is that the assumed bias be removed.

The request was timely and taken seriously by the profession of psychology. In 1969, there was actually very little research available to address the questions surrounding bias in psychological assessment. The efforts of ABPsi spurred the disciplines that develop and apply tests to create standards and conduct empirical inquiry into these issues. Today, we know a great deal about the problems of bias in psychological tests and assessments.

Many potentially legitimate objections to the use of educational and psychological tests with minorities have been raised by Black and other minority psychologists. Unfortunately, these objections are frequently stated, still, as facts, on rational rather than empirical grounds. The most frequently stated problems fall into one of the following categories (Reynolds, 2000; Reynolds et al., 1999; Reynolds & Ramsay, 2003).

15.5.1 Inappropriate Content

Black and other minority children have not been exposed to the material involved in the test questions or other stimulus materials. The tests are geared primarily toward White middle-class homes, vocabulary, knowledge, and values. As a result of inappropriate content, the tests are unsuitable for use with minority children.

15.5.2 Inappropriate Standardization Samples

Ethnic minorities are underrepresented in standardization samples used in the collection of normative reference data. As a result of the inappropriate standardization samples, the tests are unsuitable for use with minority children.

15.5.3 Examiner and Language Bias

Because most psychologists are White and speak only standard English, they may intimidate Black and other ethnic minorities and so examiner and language bias

result. They are also unable accurately to communicate with minority children—to the point of being insensitive to ethnic pronunciation of words on the test. Lower test scores for minorities, then, may reflect only this intimidation and difficulty in the communication process, not lower ability.

15.5.4 Inequitable Social Consequences

As a result of bias in educational and psychological tests, minority group members, already at a disadvantage in the educational and vocational markets because of past discrimination, are thought to be unable to learn and are disproportionately assigned to dead-end educational tracks. This represents inequitable social consequences. Labeling effects also fall under this category.

15.5.5 Measurement of Different Constructs

Related to inappropriate test content mentioned earlier, this position asserts that the tests measure different constructs when used with children from other than the middle-class culture on which the tests are largely based, and thus do not measure minority intelligence validly.

15.5.6 Differential Predictive Validity

> The hypothesis of differential validity suggests that tests measure constructs more accurately and make more valid predictions for individuals from the groups on which the tests are mainly based than for those from other groups.

Although tests may accurately predict a variety of outcomes for middle-class children, they do not predict successfully any relevant behavior for minority group members. In other words, test usage might result in valid predictions for one group, but invalid predictions in another. This is referred to as differential predictive validity. Further, there are objections to the use of the standard criteria against which tests are validated with minority cultural groups. For example, scholastic or academic attainment levels in White middle-class schools are themselves considered by a variety of Black psychologists to be biased as criteria for the validation of aptitude measures.

15.5.7 Qualitatively Distinct Aptitude and Personality

Minority and majority groups possess aptitude and personality characteristics that are qualitatively different, and as a result test development should begin with

different definitions for different groups. For example, Richardson (1993) holds that researchers have not satisfactorily settled the debate over whether intelligence tests measure general intelligence or a European cognitive style. Similarly, Helms (1992) proposed a cognitive-difference model that emphasizes differences in "European--centered" and "African-centered" values and beliefs. Helms suggests that these different styles significantly impact the way examinees respond on intelligence tests, which would then require different item sets or at least different "correct answers" from individuals of different ethnic backgrounds. Special Interest Topic 15.4 provides an introduction to a unique explanation for group differences referred to as "stereotype threat."

Special Interest Topic 15.4: Stereotype Threat: An Emerging But Controversial Explanation of Group Differences on Various Tests of Mental Abilities

Steele and Aronson in 1995 posited a unique explanation for group differences on mental test scores. They argued that such differences were created by a variable they deemed "Stereotype Threat." More recently, they defined stereotype threat as follows:

> When a negative stereotype about a group that one is part of becomes relevant, usually as an interpretation of one's behavior or an experience one is having, stereotype threat is the resulting sense that one can then be judged or treated in terms of the stereotype or that one might do something that would inadvertently confirm it (Steele, Spencer, & Aronson, 2002, p. 389).

While we find this explanation somewhat vague and lacking specificity for research purposes, in experimental research regarding mental testing outcomes, stereotype threat is most often operationalized as being given a test that is described as diagnostic of one's ability and/or being asked to report one's race prior to testing. Therefore we see two components to the threat—being told one's ability is to be judged on a test of mental ability and secondly, being asked to report one's racial identification, or at least believing it to be relevant in some way to the evaluation of examination results (although some argue either component is sufficient to achieve the effect). Stereotype threat research then goes on to argue, as one example, that if one takes a test of mental ability, but the examinee is told it is not for evaluating the test taker, but to examine the test itself and no racial identifier is requested, then racial group differences in performance on the test will disappear.

Many studies now demonstrate this stereotype effect, but some incorporate controversial statistical procedures that might confound the results by equating the two groups (i.e., erasing the group differences) on the basis of variables irrelevant to the effect of the stereotype threat. Sackett and his colleagues (2004) have discussed this methodological problem in detail (noting additional violations of the assumptions that underlie such analyses), and we find ourselves in essential agreement with their observations. Nomura et al. (2007) stated it succinctly when they noted from their own findings:

"Equalizing the performance of racial groups in most Stereotype Threat Studies is not an effect of the manipulation of Stereotype Threat elicitors (task descriptions), but is a result of a statistical manipulation (covariance)" (p. 7). Additionally, some research that has taken a thorough look at the issue using multiple statistical approaches has argued that stereotype threat may have just the opposite effect at times from what was originally proposed by Steele and Aronson (e.g., see Nomura et al., 2007). That is, it may enhance the performance of the majority group as opposed to denigrating the performance of the minority.

We are also bothered by the theoretical vagaries of the actual mechanism by which stereotype threat might operate as a practical matter. Steele and Aronson essentially argue that it is a process of response inhibition; that is, when an individual encounters a circumstance, event, or activity in which a stereotype of a group to which the person belongs becomes salient, anxiety or concerns about being judged according to that stereotype arise and inhibit performance. Anxiety is not named specifically as the culprit by many stereotype threat researchers, but it seems the most likely moderator of the proclaimed effect. While the well-known inverted U-shaped anxiety-performance curve seems real enough, can this phenomenon really account for group differences in mental test scores? So far, we view the findings of racial equalization due to the neutralization of the so-called stereotype effect as a statistical artifact, but the concept remains interesting, is not yet fully understood, and we may indeed be proven wrong!

Some good readings on this issue for follow up include the following works:

Nomura, J. M., Stinnett, T., Castro, F., Atkins, M., Beason, S., Linden, S., … Wiechmann, K. (2007, March). *Effects of stereotype threat on cognitive performance of African Americans*. Paper presented to the annual meeting of the National Association of School Psychologists, New York.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African-American differences on cognitive tests. *American Psychologist, 59*(1), 7–13.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 379–440). New York, NY: Academic Press.

The early actions of the ABPsi were most instrumental in bringing forward these objections into greater public and professional awareness and subsequently

prompted a considerable amount of research. When the objections were first raised, very little data existed to answer these charges. Contrary to the situation decades ago when the current controversy began, research now exists that examines many of these concerns and does so in great detail. Test developers and publishers also routinely examine tests for potentially biasing factors as well, prior to making tests commercially available.

> The early actions of the ABPsi brought these issues into public and professional awareness, which subsequently promoted a considerable amount of research.

15.6 The Problem of Definition in Test Bias Research: Differential Validity

Arriving at a consensual definition of test bias has produced considerable as yet unresolved debate among many measurement professionals. Although the resulting debate has generated a number of models from which to examine bias, these models usually focus on the decision-making system and not on the test itself. The concept of test bias per se then comes down to a question of the validity of the proposed interpretation of performance on a test and the estimation of that performance level, that is, the test score. Test bias refers to systematic error in the estimation of some "true" value for a group of individuals, due to construct underrepresentation or due to construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers (AERA et al., 2014). As we noted previously, differential validity is present when a test measures or estimates a construct differently for one group than for another.

As discussed in previous chapters, evidence for the validity of test score interpretations can come from sources both internal and external to the test. Bias in a test may be found to exist in any or all of these categories of validity evidence. Prior to examining the evidence on the CTBH, the concept of culture-free testing and the definition of mean differences in test scores as test bias merit attention (Special Interest Topic 15.5).

Special Interest Topic 15.5: Why Is the Cultural Test Bias Hypothesis So Robust in the Profession: And Why Do So Many Hold the Mean Differences Equals Bias Ideology?

Since the late 1960s, a substantial body of content and methodological research on bias has been conducted. Much of this research has been conducted by psychometricians and published in major psychometric journals not often read by those in other psychological specialties. However, much of what has been learned is summarized in book chapters and entire books easily accessible to mainstream psychologists and some of the empirical and methodological research has appeared in the most widely subscribed journals of the American Psychological Association. Nevertheless, certain myths persist

in the writings and actions of many professional psychologists who are either unaware of this research or choose to ignore it. Below are some thoughts of this seeming conundrum.

With regard to the media, Herrnstein (1982) provided an account of his own media encounters that leads one to believe that bias in the media is responsible for their selective reporting (see also Brown et al., 1999, for additional examples). Perhaps this is the case, but with specific regard to race and ethnic differences on measures of IQ, aptitude, and achievement, a broader explanation—and one that captures the biases of psychologists—seems required. Whatever this explanation may be, it is likely related to the phenomenon that leads some in our profession to believe in miraculous cures for intellectual disability and the dramatic resistance to the discrediting of the Milwaukee Project (Reynolds, 1987). (The Milwaukee Project was discussed in Chap. 9. In summary, the project's initial results supported the hypothesis that early, intensive interventions with children at risk for intellectual disability could dramatically increase intelligence and future academic achievement. However, subsequent research found little or no support for the hypothesis that early interventions could result in lasting changes in IQ or achievement.)

Particularly for health providers, but for most of the lay public as well, it is our concern, our hope, and our belief in our fellow humans that leads to ready acceptance of the CTBH and to the idea that any mean difference in scores or performance levels on psychological tests confirms that tests are biased. We want everyone to be created equal not just in the sense of worth as a human being as acknowledged in our Constitution, but in the sense of level of aptitude or ability. We find it anathema that ethnic differences in aptitude or ability might be real; we simply do not want it to be so. So, we search for reasons for why these differences are not true. The CTBH seems far more palatable than the alternative, as it argues that racial and ethnic group differences on mental tests result from problems with the tests themselves—tests also being something for which we all have some, though varying degrees of dislike anyway. The emotional and political appeal of the hypothesis is strong but dangerous. It is also the appeal of the egalitarian fallacy.

Some who do read the psychometric research dismiss it in favor of political arguments. Gould (1995, 1996) has acknowledged that tests are not statistically biased and do not show differential predictive validity. He argues, however, that defining cultural bias statistically is confusing: The public is concerned not with statistical bias, but with whether minority-White IQ differences occur because society treats ethnic minorities unfairly. That is, the public considers tests biased if they record biases originating elsewhere in society (Gould, 1995). In this context we interpret the tests as the messengers—the Gould approach is the "kill the messenger" approach and does not lead to solutions; rather, it leads to ignorance.

A related issue that is also likely involved in the profession's reluctance to abandon the CTBH is a failure to separate the CTBH from questions of etiology. Data reflecting ethnic differences on aptitude measures have been interpreted as supporting the hypothesis of genetic differences in intelligence and implicating one group as superior to another. Such interpretations understandably call for an emotional response and are not defensible from a scientific perspective.

The task of science and rational inquiry is to understand the source of these differences, to pit alternative theories boldly against one another, to analyze and consider our data and its complexities again and again, and to do so without the emotional overpull of our compassion and our beliefs. Particularly with regard to such sensitive and polemic topics as racial and ethnic differences on mental tests, we must stay especially close to our empirical research, perhaps adopting an old but articulate rubric, the one with which we opened this chapter: In God we trust; all others must have data.

15.7 Cultural Loading, Cultural Bias, and Culture-Free Tests

Cultural loading and cultural bias are not synonymous terms, though the concepts are frequently confused even in the professional literature. A test or test item can be culturally loaded without being culturally biased. Cultural loading refers to the degree of cultural specificity present in the test or individual items of the test. Certainly, the greater the cultural specificity of a test item, the greater the likelihood of the item being biased when used with individuals from other cultures. Virtually all tests in current use are bound in some way by their cultural specificity. Culture loading must be viewed on a continuum from general (defining the culture in a broad, liberal sense) to specific (defining the culture in narrow, highly distinctive terms).

> Cultural loading refers to the degree of cultural specificity present in the test or individual items of the test.

A number of attempts have been made to develop a culture-free (sometimes referred to as culture fair) intelligence test. However, ***culture-free tests*** are generally inadequate from a statistical or psychometric perspective (e.g., Anastasi & Urbina, 1997). It may be that because intelligence is often defined in large part on the basis of behavior judged to be of value to the survival and improvement of the culture and the individuals within that culture, a truly culture-free test would be a poor predictor of intelligent behavior within the cultural setting. Once a test has been developed within a culture (a culture-loaded test) its generalizability to other cultures or subcultures within the dominant societal framework becomes a matter for empirical investigation.

15.8 Inappropriate Indicators of Bias: Mean Differences and Equivalent Distributions

Differences in mean levels of performance on cognitive tasks between two groups historically (and mistakenly) are believed to constitute test bias by a number of writers (e.g., Alley & Foster, 1978; Chinn, 1979; Hilliard, 1979). Those who support mean differences as an indication of test bias state correctly that there is no valid a priori scientific reason to believe that intellectual or other cognitive performance levels should differ across race. It is the inference that tests demonstrating such differences are inherently biased that is faulty. Just as there is no a priori basis for deciding that differences exist, there is no a priori basis for deciding that differences do not exist. From the standpoint of the objective methods of science, a priori or premature acceptance of either hypothesis (differences exist versus differences do not exist) is untenable. As stated in the *Standards* (AERA et al., 2014):

> Certainly, most testing professionals agree that group differences in testing outcomes should trigger heightened scrutiny for possible sources of test bias … However, group differences in outcomes do not in themselves indicate that a testing application is biased or unfair. (p. 54)

The mean difference definition of test bias is the most uniformly rejected of all definitions of test bias by psychometricians involved in investigating the problems of bias in assessment.

Some adherents to the "mean differences as bias" position also require that the distribution of test scores in each population or subgroup be identical prior to assuming that the test is nonbiased, regardless of its validity. Portraying a test as biased regardless of its purpose or the validity of its interpretations conveys an inadequate understanding of the psychometric construct and issues of bias. The mean difference definition of test bias is the most uniformly rejected of all definitions of test bias by psychometricians involved in investigating the problems of bias in assessment (e.g., Camilli & Shepard, 1994; Cleary et al., 1975; Cole & Moss, 1989; Hunter, Schmidt, & Rauschenberger, 1984; Reynolds, 1982, 1995, 2000).

Jensen (1980) discusses the mean differences as bias definition in terms of the egalitarian fallacy. The egalitarian fallacy contends that all human populations are in fact identical on all mental traits or abilities. Any differences with regard to any aspect of the distribution of mental test scores indicate that something is wrong with the test itself. As Jensen points out, such an assumption is scientifically unwarranted. There are simply too many examples of specific abilities and even sensory capacities that have been shown to differ unmistakably across human populations. The result of the egalitarian assumption then is to remove the investigation of population differences in ability from the realm of scientific inquiry, an unacceptable course of action (Reynolds, 1980).

The belief of many people in the mean differences as bias definition is quite likely related to the nature-nurture controversy at some level. Certainly data

reflecting racial differences on various aptitude measures have been interpreted to indicate support for a hypothesis of genetic differences in intelligence and implicating one group as superior to another. Such interpretations understandably call for a strong emotional response and are not defensible from a scientific perspective. Although IQ and other aptitude test score differences undoubtedly occur, the differences do not indicate deficits or superiority by any group, especially in relation to the personal worth of any individual member of a given group or culture.

15.9 Bias in Test Content

Bias in the content of psychological and educational tests has been a popular topic of critics of testing.

Bias in the content of educational and psychological tests has been a popular topic of critics of testing. These criticisms typically take the form of reviewing the items, comparing them to the critics' views of minority and majority cultural environments, and then singling out specific items as biased or unfair because:

- The items ask for information that minority or disadvantaged individuals have not had equal opportunity to learn.
- The items require the child to use information in arriving at an answer that minority or disadvantaged individuals have not had equal opportunity to learn.
- The scoring of the items is improper, unfairly penalizing the minority child because the test author has a Caucasian middle-class orientation that is reflected in the scoring criterion. Thus minority children do not receive credit for answers that may be correct within their own cultures but do not conform to Anglocentric expectations—this occurs in personality tests wherein minorities may respond to various questions in ways seen as adaptive in their own subculture but as indicative of psychopathology in the mind of the test developer.
- The wording of the questions is unfamiliar to minorities and even though they may "know" the correct answer they are unable to respond because they do not understand the question.

These problems with test items cause the items to be more difficult than they should actually be when used to assess minority individuals. This, of course, results in lower test scores for minorities, a well-documented finding. Are these criticisms of test items accurate? Do problems such as these account for minority-majority group score differences on mental tests? These are questions for empirical resolution rather than armchair speculation, which is certainly abundant in the evaluation of test bias. Empirical evaluation first requires a working definition. We will define a ***biased test item*** as follows:

> An item is considered to be biased when it is demonstrated to be significantly more difficult for one group than another item measuring the same ability or construct when the overall level of performance on the construct is held constant.

There are two concepts of special importance in this definition. First, the group of items must be unidimensional; that is, they must all be measuring the same factor or dimension of aptitude or personality. Second, the items identified as biased must be differentially more difficult for one group than another. The definition allows for score differences between groups of unequal standing on the dimension in question but requires that the difference be reflected on all items in the test and in an equivalent fashion across items. A number of empirical techniques are available to locate deviant test items under this definition. Many of these techniques are based on item--response theory (IRT) and designed to detect differential item functioning, or DIF. The relative merits of each method are the subject of substantial debate, but in actual practice, each method has led to similar general conclusions, though the specific findings of each method often differ.

With multiple-choice tests, another level of complexity can easily be added to the examination of content bias. With a multiple-choice question, typically three or four distracters are given in addition to the correct response. Distracters may be examined for their attractiveness (the relative frequency with which they are chosen) across groups. When distracters are found to be disproportionately attractive for members of any particular group, the item may be defined as biased.

Research that includes thousands of subjects and more than 100 published studies consistently finds very little bias in tests at the level of the individual item. Although some biased items are nearly always found, they seldom account for more than 2–5% of the variance in performance and often, for every item favoring one group, there is an item favoring the other group.

> Content bias in well-prepared standardized tests is irregular in its occurrence, and no common characteristics of items that are found to be biased can be ascertained by expert judges.

Earlier in the study of item bias, it was hoped that the empirical analysis of tests at the item level would result in the identification of a category of items having similar content as biased and that such items could then be avoided in future test development (Flaugher, 1978). Very little similarity among items determined to be biased has been found. No one has been able to identify those characteristics of an item that cause the item to be biased. It does seem that poorly written, sloppy, and ambiguous items tend to be identified as biased with greater frequency than those items typically encountered in a well-constructed, standardized instrument.

A common practice of test developers seeking to eliminate "bias" from their newly developed educational and psychological tests has been to arrange for a panel of expert minority group members to review all proposed test items. Any item identified as "culturally biased" by the panel of experts is then expurgated from the instrument. Because, as previously noted, no detectable pattern or common characteristic of individual items statistically shown to be biased has been observed (given reasonable care at the item writing stage), it seems reasonable to question the armchair or expert minority panel approach to determining biased items. Several researchers, using a variety of psychological and educational tests, have identified items as being disproportionately more difficult for minority group members than

for members of the majority culture and subsequently compared their results with a panel of expert judges. Studies by Jensen (1976) and Sandoval and Mille (1979) are representative of the methodology and results of this line of inquiry.

After identifying the eight most racially discriminating and eight least racially discriminating items on the Wonderlic Personnel Test, Jensen (1976) asked panels of five Black psychologists and five Caucasian psychologists to sort out the eight most and eight least discriminating items when only these 16 items were presented to them. The judges sorted the items at a no better than chance level. Sandoval and Mille (1979) conducted a somewhat more extensive analysis using items from the WISC-R. These two researchers had 38 Black, 22 Hispanic, and 40 White university students from Spanish, history, and education classes identify items from the WISC-R that are more difficult for a minority child than a White child and items that are equally difficult for each group. A total of 45 WISC-R items was presented to each judge; these items included the 15 most difficult items for Blacks as compared to Whites, the 15 most difficult items for Hispanics as compared to Whites, and the 15 items showing the most nearly identical difficulty indexes for minority and White children. The judges were asked to read each question and determine whether they thought the item was (1) easier for minority than White children, (2) easier for White than minority children, or (3) of equal difficulty for White and minority children. Sandoval and Mille's (1979) results indicated that the judges were not able to differentiate between items that were more difficult for minorities and items that were of equal difficulty across groups. The effects of the judges' ethnic backgrounds on the accuracy of their item bias judgments were also considered. Minority and nonminority judges did not differ in their ability to identify accurately biased items nor did they differ with regard to the type of incorrect identification they tended to make. Sandoval and Mille's (1979) two major conclusions were that "1) judges are not able to detect items which are more difficult for a minority child than an Anglo child, and 2) the ethnic background of the judge makes no difference in accuracy of item selection for minority children" (p. 6). Research since that time has continued to produce similar results: minority judges seldom exceed chance expectations in designating biased versus nonbiased test items in aptitude and in personality domains. Even without empirical support for its validity, the use of expert panels of minorities continues but for a different purpose. Members of various ethnic, religious, or other groups that have a cultural system in some way unique may well be able to identify items that contain material that is offensive, and the elimination of such items is proper.

15.9.1 How Test Publishers Commonly Identify Biased Items

Today's most recommended method for detecting ***item bias*** arises from applications of item-response theory (IRT), followed by a thoughtful, logical analysis of item content (Reynolds, 2000). The goal in the use of these methods is to determine the degree of differential item functioning (DIF), that is, whether items function differently across groups, as indicated by model parameters associated with the items. Embretson and Reise (2000) present an excellent overview of the theory and

applications of item-response theory, including a highly readable chapter on detecting DIF. Statistically significant DIF, coupled with a logical analysis of item content that suggests the item may measure construct-irrelevant differences across groups, provides a basis for rejecting items from tests.

IRT is concerned fundamentally with creating a mathematical model of item difficulty— or more technically, the probability of occurrence of a particular response to a test item as a function of an examinee's relative position on a latent trait. Such models specify various parameters that describe the behavior of the item within the model; most IRT models include one, two, or three parameters, which may be graphically represented in an item characteristic curve (ICC). The three parameters in the three-parameter (3P) model are (a) discrimination power of the item, or slope of the ICC, (b) item difficulty, located at the point on the difficulty level of the latent trait at which the examinee has a 50% chance of correctly answering the item, and (c) guessing parameter. Figure 15.1 demonstrates an ICC that uses a one-parameter (also known as the Rasch Model after its originator) unidimensional model which is widely used in aptitude testing. Its appropriateness depends on the context so other ICC models may be more appropriate, particularly for multiple-choice items. The greater complexity that can be modeled with the 3P model comes with a price: one generally requires a much larger sample to develop a valid and reliable 3P model. Two computer programs widely used to estimate item and latent parameters are LOGIST and BILOG, which use the joint maximum likelihood (JML) and marginal maximum likelihood (MML) methods, respectively.

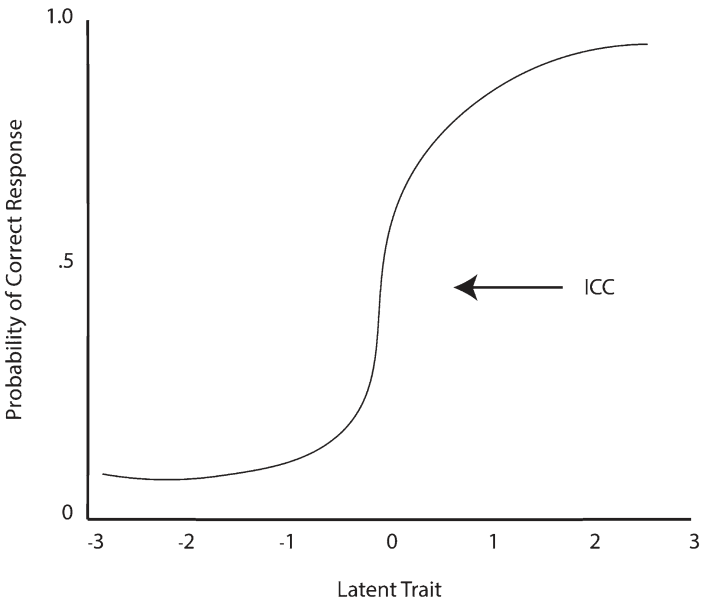

Fig. 15.1 An example of an item characteristic curve or ICC

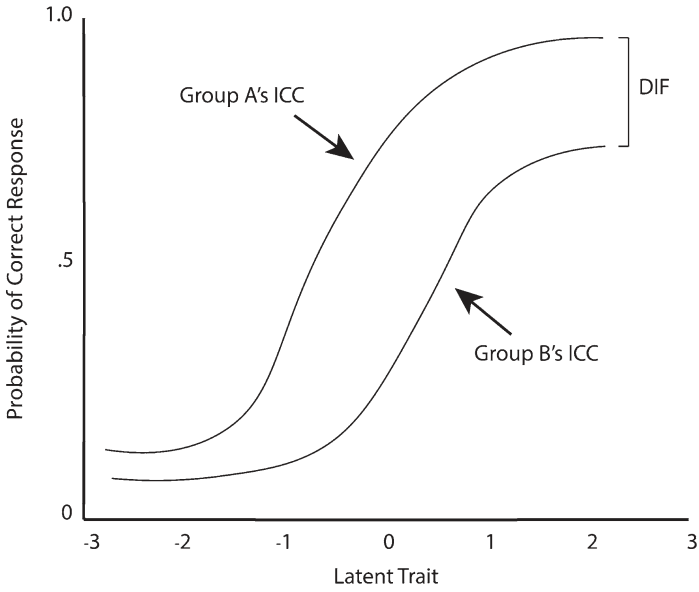

Fig. 15.2 A visual representation of DIF is the region between group A's and group B's item characteristic curves (ICCs)

In using IRT to determine DIF, one compares the ICCs of two different groups, yielding a DIF "index." Various statistical methods have been developed for measuring the gaps between ICCs across groups of examinees. Figure 15.2 demonstrates this DIF "gap" across two hypothetical groups.

Another method used by test publishers is a partial correlation approach. Using partial correlations, one may test for differences between groups on the degree to which there exists significant or meaningful variation in observed scores on individual test items not attributable to the total test score. It provides a simpler method than either the ICC or IRT models, and it is readily accessible through major statistical programs. In this method, the partial correlation between item score and the nominal variable of interest (e.g., sex) is calculated, partialling the correlation between total test score and the nominal variable. This method essentially holds the total score constant across the groups, and the resulting differences may be used to identify problematic items if it is significant and particularly if meaningful. This latter determination commonly based on effect size, which is easily obtained by squaring the partial *r* value. Reynolds, Willson, and Chatman (1984) provide more information on the development of this method. The main risk of the use of this method is that it may over-identify differences between groups, so it is necessary to calculate experiment-wise error rates. However, this also makes the partial correlation method more sensitive to potentially biased items.

Reynolds and Kamphaus (2003) used the partial correlation method to detect potentially bias items in their development of the first edition of the Reynolds

Intellectual Assessment Scales (RIAS). They computed the partial *r* of the item--subtest total score, partialling the total score correlation with each nominal variable of interest (gender and ethnicity) one variable at a time and separately by age group. One advantage of the partial correlation in such studies is that it can be used successfully with much smaller sample sizes than the ICC and most other techniques. Thus, analyses can be run at smaller age intervals and any developmental interaction can be detected more readily. The partial *r* and subsequently its effect size stabilize at smaller sample sizes compared to the IRT approach.

From a large number of studies employing a wide range of methodologies, a relatively clear picture emerges. Content bias in well-prepared standardized tests is irregular in its occurrence, and no common characteristics of items that are found to be biased can be ascertained by expert judges (minority or nonminority). The variance in group score differences on mental tests associated with ethnic group membership when content bias has been found is relatively small (typically ranging from 2% to 5%). Although the search for common biased item characteristics will continue, cultural bias in aptitude tests has found no consistent empirical support in a large number of actuarial studies contrasting the performance of a variety of ethnic and gender groups on items of the most widely employed intelligence scales in the United States. Most major test publishing companies do an adequate job of reviewing their assessments for the presence of content bias. Nevertheless, certain standardized tests have not been examined for the presence of content bias, and research with these tests should continue regarding potential content bias with different ethnic groups (Reynolds & Ramsay, 2003).

15.10 Bias in Other Internal Features of Tests

There is no single method for the accurate determination of the degree to which educational and psychological tests measure a distinct construct. The defining of bias in construct measurement, i.e., ***construct bias***, then requires a general statement that can be researched from a variety of viewpoints with a broad range of methodology. The following rather parsimonious definition is proffered:

> Bias exists in regard to construct measurement when a test is shown to measure different hypothetical traits (psychological constructs) for one group than another or to measure the same trait but with differing degrees of accuracy. (Reynolds, 1982)

As is befitting the concept of construct measurement, many different methods have been employed to examine existing psychological tests and batteries of tests for potential bias. One of the more popular and necessary empirical approaches to investigating construct measurement is factor analysis. Factor analysis, as a procedure, identifies clusters of test items or clusters of subtests of psychological or educational tests that correlate highly with one another, and less so or not at all with other subtests or items. Factor analysis allows one to determine patterns of interrelationships of performance among groups of individuals. For example, if several subtests of an intelligence scale load highly on (are members of) the same factor,

then if a group of individuals score high on one of these subtests, they would be expected to score at a high level on other subtests that load highly on that factor. Psychometricians attempt to determine through a review of the test content and correlates of performance on the factor in question what psychological trait underlies performance; or, in a more hypothesis testing approach, they will make predictions concerning the pattern of factor loadings. Hilliard (1979), one of the more vocal critics of IQ tests on the basis of cultural bias, pointed out early in test bias research that one of the potential ways of studying bias involves the comparison of factor analytic results of test studies across race.

If the IQ test is a valid and reliable test of "innate" ability or abilities, then the factors which emerge on a given test should be the same from one population to another, since "intelligence" is asserted to be a set of mental processes. Therefore, while the configuration of scores of a particular group on the factor profile would be expected to differ, logic would dictate that the factors themselves would remain the same (p. 53).

Although not agreeing that identical factor analyses of an instrument speak to the "innateness" of the abilities being measured, consistent factor analytic results across populations do provide strong evidence that whatever is being measured by the instrument is being measured in the same manner and is in fact the same construct within each group. The information derived from comparative factor analysis across populations is directly relevant to the use of educational and psychological tests in diagnosis and other decision-making functions. Psychologists, in order to make consistent interpretations of test score data, must be certain that the test(s) measures the same variable across populations.

A number of studies of factorial similarity of tests' latent structures have appeared over the past three decades, dealing with a number of different tasks. These studies have for the most part focused on aptitude or intelligence tests, the most controversial of all techniques of measurement. Numerous studies of the similarity of factor analysis outcomes for children of different ethnic groups, across gender, and even diagnostic groupings have been reported over the past 30 years. Results reported are highly consistent in revealing that the internal structure of most standardized tests varies quite little across groups. Comparisons of the factor structure of the Wechsler Intelligence Scales (e.g., various editions of the WISC and WAIS) and the Reynolds Intellectual Assessment Scales (Reynolds & Kamphaus, 2003) in particular and other intelligence tests find the tests to be highly factorially similar across gender and ethnicity for Blacks, Whites, and Hispanics. The structure of ability tests for other groups has been researched less extensively, but evidence thus far with Chinese, Japanese, and Native Americans does not show substantially different factor structures for these groups.

As is appropriate for studies of construct measurement, comparative factor analysis has not been the only method of determining whether bias exists. Another method of investigation involves the comparison of internal-consistency reliability estimates across groups. As described in Chap. 4, internal-consistency reliability is determined by the degree to which the items are all measuring a similar construct. The internal-consistency reliability coefficient reflects the accuracy of measurement

of the construct. To be unbiased with regard to construct validity, internal-consistency estimates should be approximately equal across race. This characteristic of tests has been investigated for a number of popular aptitude tests for Blacks, Whites, and Hispanics with results similar to those already noted.

15.10.1 How Test Publishers Commonly Identify Bias in Construct Measurement

Factor analysis across groups is the most common method in use by various commercial test developers to assess for bias in construct measurement. However, many other methods of comparing construct measurement across groups have been used to investigate bias in tests. These methods include the correlation of raw scores with age, comparison of item-total correlations across groups, comparisons of alternate form and test-retest correlations, evaluation of kinship correlation and differences, and others (see Reynolds, 2002, for a discussion of these methods). A more recently proposed method for assessing test bias is comparative item selection (Reynolds, 1998). This method involves the use of the same method of selecting items for inclusion in a test repeated across the groups of interest; one is free to use item selection methods based on either classical test theory or IRT. Unbiased tests will generally obtain about a 90% rate of overlap between selected items. The technique will yield substantially lower rate of overlap with biased tests, as well as tests with poor item reliabilities. This method also requires large samples for stable results. Reynolds (1998) provides a full discussion of this approach and demonstrates its application to several personality measures. The general results of research with all of these methods have been supportive of the consistency of construct measurement of tests across ethnicity and gender.

Construct measurement of a large number of popular psychometric assessment instruments has been investigated across ethnicity and gender with a divergent set of methodologies. No consistent evidence of bias in construct measurement has been found in the many prominent standardized tests investigated. This leads to the conclusion that these psychological tests function in essentially the same manner across ethnicity and gender, the test materials are perceived and reacted to in a similar manner, and the tests are measuring the same construct with equivalent accuracy for Blacks, Whites, Hispanic, and other American minorities for both sexes. Differential validity or single-group validity has not been found and likely is not an existing phenomenon with regard to well-constructed standardized psychological and educational tests. These tests appear to be reasonably unbiased for the groups investigated, and mean score differences do not appear to be an artifact of test bias (Reynolds & Ramsay, 2003).

No consistent evidence of bias in construct measurement has been found in the many prominent standardized tests investigated.

15.11 Bias in Prediction and in Relation to Variables External to the Test

Internal analyses of bias (such as with item content and construct measurement) are less confounded than analyses of bias in prediction due to the potential problems of bias in the criterion measure. Prediction is also strongly influenced by the reliability of criterion measures, which frequently is poor. (The degree of relation between a predictor and a criterion is restricted as a function of the square root of the product of the reliabilities of the two variables.) Arriving at a consensual definition of bias in prediction is also a difficult task. Yet, from the standpoint of the traditional practical applications of aptitude and intelligence tests in forecasting probabilities of future performance levels, prediction is the most crucial use of test scores to examine. Looking directly at bias as a characteristic of a test and not a selection model, Cleary et al.'s (1975) definition of test fairness, as restated here in modern times, is a clear direct statement of test bias with regard to ***prediction bias***:

From the standpoint of traditional practical applications on aptitude and intelligence tests in forecasting probabilities of future performance levels, prediction is the most crucial use of test scores to examine.

> A test is considered biased with respect to prediction when the inference drawn from the test score is not made with the smallest feasible random error or if there is constant error in an inference or prediction as a function of membership in a particular group. (Reynolds, 1982, p. 201)

The evaluation of bias in prediction under the Cleary et al. (1975) definition (known as the regression definition) is quite straightforward. With simple regressions, predictions take the form $Y = aX + b$ where a is the constant and b is the regression coefficient. When this equation is graphed (forming a regression line), a is the Y-intercept and b the slope of the regression line. Given our definition of bias in prediction validity, nonbias requires errors in prediction to be independent of group membership, and the regression line formed for any pair of variables must be the same for each group for whom predictions are to be made. Whenever the slope or the intercept differs significantly across groups, there is bias in prediction if one attempts to use a regression equation based on the combined groups. When the regression equations for two (or more) groups are equivalent, prediction is the same for those groups. This condition is referred to variously as homogeneity of regression across groups, simultaneous regression, or fairness in prediction. ***Homogeneity of regression*** is illustrated in Fig. 15.3, in which the regression line shown is equally appropriate for making predictions for all groups. Whenever homogeneity of regression across groups does not occur, then separate regression equations should be used for each group concerned.

When the regression equations are the same for two or more groups, prediction is the same for those groups.

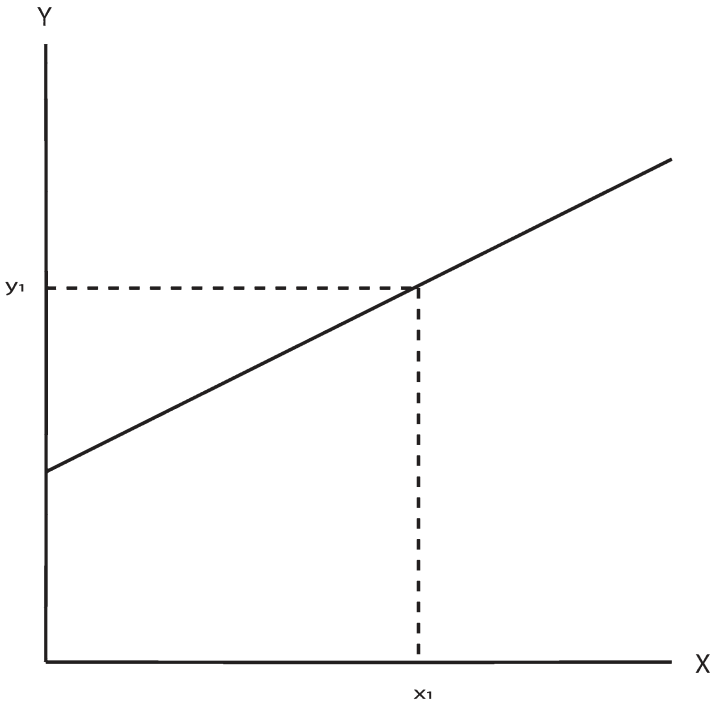

Fig. 15.3 Equal slopes and intercepts. Note: Equal slopes and intercepts result in homogeneity of regression where the regression lines for different groups are the same

In actual clinical practice, regression equations are seldom generated for the prediction of future performance. Rather, some arbitrary or perhaps statistically derived cutoff score is determined, below which failure is predicted. For school performance, a score of 2 or more standard deviations below the test mean is used to infer a high probability of failure in the regular classroom if special assistance is not provided for the student in question. Essentially then, clinicians are establishing prediction equations about mental aptitude that are assumed to be equivalent across race, sex, and so on. Although these mental equations cannot be readily tested across groups, the actual form of criterion prediction can be compared across groups in several ways. Errors in prediction must be independent of group membership. If regression equations are equal, this condition is met. To test the hypothesis of simultaneous regression, regression slopes and regression intercepts must both be compared.

When homogeneity of regression does not occur, three basic conditions can result: (1) Intercept constants differ, (2) regression coefficients (slopes) differ, or (3) slopes and intercepts differ. These conditions are illustrated in Figs. 15.4, 15.5, and 15.6, respectively.

When intercept constants differ, the resulting bias in prediction is constant across the range of scores. That is, regardless of the level of performance on the

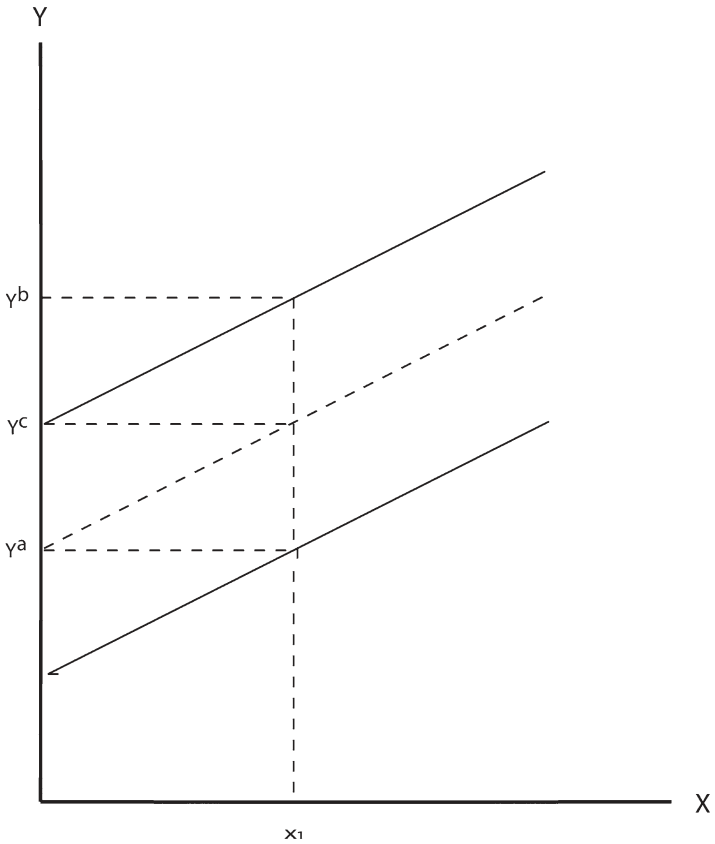

Fig. 15.4 Equal slopes with differing intercepts. Note: Equal slopes with differing intercepts result in parallel regression lines that produce a constant bias in prediction

independent variable, the direction and degree of error in the estimation of the criterion (systematic over- or underprediction) will remain the same. When regression coefficients differ and intercepts are equivalent, the direction of the bias in prediction will remain constant, but the amount of error in prediction will vary directly as a function of the distance of the score on the independent variable from the origin. With regression coefficient differences, then, the higher the score on the predictor variable, the greater the error of prediction for the criterion. When both slopes and intercepts differ, the situation becomes even more complex. Both the degree of error in prediction and the direction of the "bias" will vary as a function of level of performance on the independent variable.

A considerable body of literature has developed over the last 40 years regarding differential prediction of tests across ethnicity for employment selection, college admissions, and school or academic performance generally. In an impressive review of 866 Black-White prediction comparisons from 39 studies of test bias in personnel selection, Hunter, Schmidt, and Hunter (1979) concluded that there was no

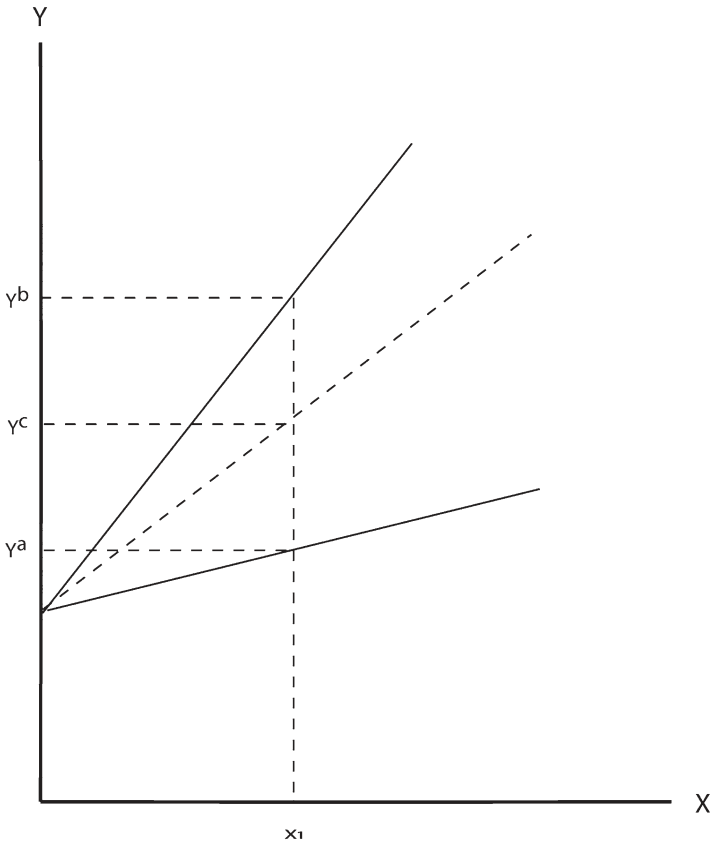

Fig. 15.5 Equal intercepts and differing slopes. Note: Equal intercepts and differing slopes result in nonparallel regression lines, with the degree of bias depending on the distance of the individual's score from the origin

evidence to substantiate hypotheses of differential or single-group validity with regard to the prediction of the job performance across race for Blacks and Whites. A similar conclusion has been reached by other independent researchers (e.g., Reynolds, 1995). A number of studies have also focused on differential validity of the Scholastic Aptitude Test (SAT) in the prediction of college performance (typically measured by grade point average). In general, these studies have found either no difference in the prediction of criterion performance for Blacks and Whites or a bias (underprediction of the criterion) against Whites. When bias against Whites has been found, the differences between actual and predicted criterion scores, while statistically significant, have generally been quite small.

Studies investigating bias in the prediction of future school performance based on IQ tests for children have covered a variety of populations including normal as well as referred children; high-poverty, inner-city children; rural Black; and Native American groups. Studies of preschool as well as school-age children have been

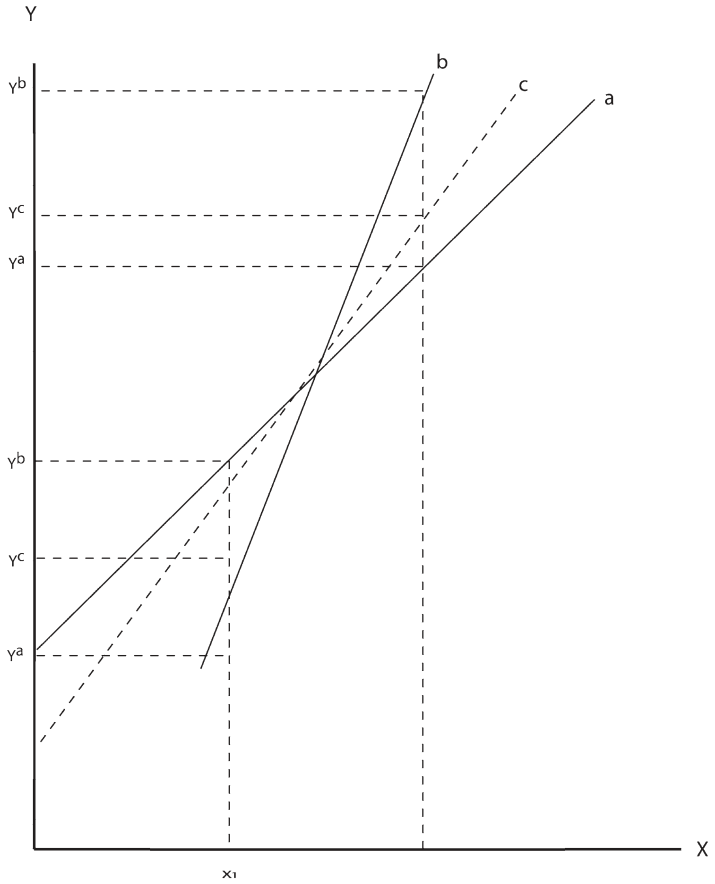

Fig. 15.6 Differing Slopes and Intercepts. Note: Differing slopes and intercepts result in a complex situation where the amount and the direction of the bias are a function of the distance of an individual's score from the origin

carried out. Almost without exception, those studies have produced results that can be adequately depicted by Fig. 15.1, that is, equivalent prediction for all groups. When this has not been found, intercepts have generally differed resulting in a constant bias in prediction. Yet, the resulting bias has not been in the popularly conceived direction. The bias identified has tended to overpredict how well minority children will perform in academic areas and to underpredict how well White children will perform. Reynolds (1995) provides a thorough review of studies investigating the prediction of school performance in children.

With regard to bias in prediction, the empirical evidence suggests conclusions similar to those regarding bias in test content and other internal characteristics. There is no strong evidence to support contentions of differential or single-group validity. Bias occurs infrequently and with no apparently observable pattern, except with regard to instruments of poor reliability and high specificity of test content.

When bias occurs, it usually takes the form of small overpredictions for low SES, disadvantaged ethnic minority children, or other low-scoring groups. These overpredictions are unlikely to account for adverse placement or diagnosis in these groups (Reynolds & Ramsay, 2003).

> Single-group or differential validity has not been found and likely is not an existing phenomenon with regard to well--constructed standardized psychological tests.

15.11.1 How Test Publishers Commonly Identify Bias in Prediction

Commercial test developers seldom demonstrate the presence or absence of bias in prediction prior to test publication. Unfortunately, the economics of the test development industry as well as that for researchers developing tools for specific research projects prohibit such desirable work prepublication. Most such work occurs post--publication and by independent researchers with interests in such questions.

15.12 Summary

A considerable body of literature currently exists failing to substantiate cultural bias against native-born American ethnic minorities with regard to the use of well--constructed, adequately standardized intelligence and aptitude tests. With respect to personality scales, the evidence is promising yet far more preliminary and thus considerably less conclusive. Despite the existing evidence, we do not expect the furor over the CTBH to be resolved soon. Bias in psychological testing will remain a polemic issue for some time. Psychologists and educators will need to keep abreast of new findings in the area. As new techniques and better methodology are developed and more specific populations examined, the findings of bias now seen as random and infrequent may become better understood and seen to indeed display a correctable pattern.

In the meantime, however, one cannot ethically fall prey to the sociopoliticolegal Zeitgeist of the times and infer bias where none exists (see Special Interest Topic 15.5 for further thoughts on this issue). Psychologists and educators cannot justifiably ignore the fact that low IQ, ethnic, disadvantaged children are just as likely to fail academically as are their low IQ, White, middle-class counterparts. Black adolescents with deviant personality scale scores and who exhibit aggressive behavior need treatment environments as much as their White peers with deviant personality scores and aggressive behaviors. The potential outcome for score interpretation (e.g., therapy versus prison, special education versus regular education) cannot dictate the psychological meaning of test performance. We must practice intelligent testing (Kaufman, 1994). We must remember that it is the purpose of the assessment process to beat the prediction made by the test, to provide insight into hypotheses

for environmental interventions that prevent the predicted failure or subvert the occurrence of future maladaptive behavior.

Continued sensitivity by test developers to issues of bias is also necessary so that appropriate checks for bias are performed prior to test publication. Progress is being made in all of these areas. However, we must hold to the data even if we do not like them. At present, only scattered and inconsistent evidence for bias exists. The few findings of bias do suggest two guidelines to follow in order to ensure nonbiased assessment: (1) Assessment should be conducted with the most reliable instrumentation available, and (2) multiple abilities should be assessed. In other words, educators and psychologists need to view multiple sources of accurately derived data prior to making decisions concerning individuals. One hopes that this is what has actually been occurring in the practice of assessment, although one continues to hear isolated stories of grossly incompetent placement decisions being made. This is not to say educators or psychologists should be blind to an individual's cultural or environmental background. Information concerning the home, community, and school environment must all be evaluated in individual decisions. As we noted, it is the purpose of the assessment process to beat the prediction and to provide insight into hypotheses for environmental interventions that prevent the predicted failure.

Without question, scholars have not conducted all the research that needs to be done to test the CTBH and its alternatives. A number and variety of criteria need to be explored further before the question of bias is empirically resolved. Many different achievement tests and teacher-made, classroom-specific tests need to be employed in future studies of predictive bias. The entire area of differential validity of tests in the affective domain is in need of greater exploration. A variety of views toward bias have been expressed in many sources; many with differing opinions offer scholarly, nonpolemical attempts directed toward a resolution of the issue. Obviously, the fact that such different views are still held indicates resolution lies in the future. As far as the present situation is concerned, clearly all the evidence is not in. With regard to a resolution of bias, we believe that were a scholarly trial to be held, with a charge of cultural bias brought against mental tests, the jury would likely return the verdict other than guilty or not guilty that is allowed in British law—"not proven." Until such time as a true resolution of the issues can take place, we believe the evidence and positions taken in this chapter accurately reflect the state of our empirical knowledge concerning bias in mental tests.

References

Alley, G., & Foster, C. (1978). Nondiscriminatory testing of minority and exceptional children. *Focus on Exceptional Children, 9*, 1–14.

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

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since "Bias in Mental Testing." *School Psychology Quarterly, 14*, 208–238.

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

Chinn, P. C. (1979). The exceptional minority child: Issues and some answers. *Exceptional Children, 46*, 532–536.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30*, 15–41.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). New York, NY: Macmillan.

Dickens, W. T., & Flynn, J. R. (2006). Black Americans reduce the racial IQ gap: Evidence from standardization samples. *Psychological Science, 17*, 913–920.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. London: Taylor & Francis.

Flaugher, R. (1978). The many definitions of test bias. *American Psychologist, 33*(7), 671–679.

Gould, S. J. (1995). Curveball. In S. Fraser (Ed.), *The bell curve wars: Race, intelligence, and the future of America* (pp. 11–22). New York, NY: Basic Books.

Gould, S. J. (1996). *The mismeasure of man* (rev. ed.). New York, NY: Norton.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091–1102.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*, 1083–1101.

Herrnstein, R. J. (1982, August). IQ testing and the media. *Atlantic Monthly, 250*, 68–74.

Hilliard, A. G. (1979). Standardization and cultural bias as impediments to the scientific study and validation of "intelligence". *Journal of Research and Development in Education, 12*, 47–58.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*, 721–735.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, L. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 41–100). New York, NY: Plenum Press.

Jensen, A. R. (1976). Test bias and construct validity. *Phi Delta Kappan, 58*, 340–346.

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York, NY: Wiley.

Neisser, U., BooDoo, G., Bouchard, T., Boykin, A., Brody, N., Ceci, S., … Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Nisbett, R. E. (2009). *Intelligence and how to get it: Why schools and cultures count*. New York, NY: Norton.

Nomura, J. M., Stinnett, T., Castro, F., Atkins, M., Beason, S., Linden, S., … Wiechmann, K. (2007, March). *Effects of stereotype threat on cognitive performance of African Americans*. Paper presented to the annual meeting of the National Association of School Psychologists, New York.

Reynolds, C., Willson, V., & Chatman, S. (1984). Item bias on the 1981 revision of the Peabody Picture Vocabulary Test using a new method of detecting bias. *Journal of Psychoeducational Assessment, 2*(3), 219–224.

Reynolds, C. R. (1980). In support of "Bias in Mental Testing" and scientific inquiry. *Behavioral and Brain Sciences, 3*, 352.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178–208). New York, NY: Wiley.

Reynolds, C.R. (1983). Test bias: In God we trust; all others must have data. *Journal of Special Education, 17*, 241–260.

Reynolds, C. R. (1987). Raising intelligence: Clever Hans, Candides, and the Miracle in Milwaukee. *Journal of School Psychology, 25*, 309–312.

Reynolds, C. R. (1995). Test bias in the assessment of intelligence and personality. In D. SakJofsky & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 545–576). New York, NY: Plenum Press.

Reynolds, C. R. (1998). Fundamentals of measurement and assessment in psychology. In A. Bellack & M. Hersen (Eds.), *Comprehensive clinical psychology* (pp. 33–55). New York, NY: Elsevier.

Reynolds, C. R. (2000). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law, 6*, 144–150.
Reynolds, C. R. (2002). *Comprehensive trail-making test: Examiner's manual*. Austin, TX: Pro-Ed.
Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds intellectual assessment scales*. Lutz, FL: Psychological Assessment Resources.
Reynolds, C. R., Lowe, P. A., & Saenz, A. (1999). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (3rd ed., pp. 549–595). New York, NY: Wiley.
Reynolds, C. R., & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (pp. 67–93). New York, NY: Wiley.
Richardson, T. Q. (1993). Black cultural learning styles: Is it really a myth? *School Psychology Review, 22*(3), 562–567.
Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on group differences in cognitive ability. *Psychology, Public Policy, and Law, 11*, 235–294.
Rushton, J. P., & Jensen, A. R. (2010). Race and IQ: A theory-based review of the research in Richard Nisbett's intelligence and how to get it. *The Open Psychology Journal, 3*, 9–35.
Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for african american-white differences on cognitive tests. *American Psychologist, 59*(1), 7–13.
Sandoval, J., & Mille, M. P. W. (1979). *Accuracy judgments of WISC-R item difficulty for minority groups*. Paper presented at the annual meeting of the American Psychological Association, New York.
Schoenfeld, W. N. (1974). Notes on a bit of psychological nonsense: "Race differences in intelligence". *Psychological Record, 24*, 17–32.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811.
Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). New York, NY: Academic Press.
Suzuki, L. A., & Valencia, R. R. (1997). Race-ethnicity and measured intelligence: Educational implications. *American Psychologist, 52*, 1103–1114.
Williams, R. L. (1970). Danger: Testing and dehumanizing Black children. *Clinical Child Psychology Newsletter, 9*, 5–6.
Williams, R. L., Dotson, W., Dow, P., & Williams, W. S. (1980). The war against testing: A current status report. *Journal of Negro Education, 49*, 263–273.

Recommended Reading

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). *American Psychologist, 30*, 15–41 This is the report of a group appointed by the APA's Board of Scientific Affairs to study the use of psychological and educational tests with disadvantaged students--An early and influential article.
Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. London: Taylor & Francis An excellent overview of the theory and applications of IRT.
Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091–1102 A good article that summarizes the literature on sex differences with an emphasis on educational implications.
Neisser, U., BooDoo, G., Bouchard, T., Boykin, A., Brody, N., Ceci, S., … Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101 This report of an APA task force provides an excellent review of the research literature on intelligence.

Reynolds, C. R. (1995). Test bias in the assessment of intelligence and personality. In D. Saklofsky & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 545–573). New York, NY: Plenum Press This chapter provides a thorough review of the literature.

Reynolds, C. R. (2000). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law, 6*, 144–150 This article provides a particularly good discussion of test bias in terms of public policy issues.

Reynolds, C. R., & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (pp. 67–93). New York, NY: Wiley This chapter also provides an excellent review of the literature.

Suzuki, L. A., & Valencia, R. R. (1997). Race-ethnicity and measured intelligence: Educational implications. *American Psychologist, 52*, 1103–1114 A good discussion of the topic with special emphasis on educational implications and alternative assessment methods.