CHAPTER 4

# Bias in Psychological Assessment

## An Empirical Review and Recommendations

CECIL R. REYNOLDS AND LISA A. SUZUKI



### UNDERSTANDING BIAS IN PSYCHOLOGICAL ASSESSMENT

Few issues in psychological assessment today are as polarizing among clinicians and laypeople as the use of standardized tests with minority examinees. For clients, parents, and clinicians, the central issue is one of long-term consequences that may occur when mean test results differ from one ethnic group to another—Blacks, Hispanics, American Indians, Asian Americans, and so forth. Important concerns include, among others, that psychiatric clients may be overdiagnosed, students disproportionately placed in special classes, and applicants unfairly denied employment or college admission because of purported bias in standardized tests.

Among researchers, polarization also is common. Here, too, observed mean score differences among ethnic groups are fueling the controversy, but in a different way. Alternative explanations of these differences seem to give shape to the conflict. Reynolds (2000a, 2000b) divided the most common explanations into four categories: (1) genetic influences; (2) environmental factors involving economic, social, and educational deprivation; (3) an interactive effect of genes and environment; and (4) biased tests that systematically underrepresent minorities' true aptitudes or abilities. The last two of these explanations have drawn the most attention. Williams (1970) and Helms (1992) proposed a fifth interpretation of differences between Black and White examinees: The two groups have qualitatively different cognitive structures, which must be measured using different methods (Reynolds, 2000b).

The problem of cultural bias in mental tests has drawn controversy since the early 1900s, when Binet's first intelligence scale was published and Stern introduced procedures for testing intelligence (Binet & Simon, 1916/1973; Stern, 1914). The conflict is in no way limited to cognitive ability tests, but the so-called IQ controversy has attracted most of the public attention. A number of authors have published works on the subject that quickly became controversial (Gould, 1981; Herrnstein & Murray, 1994; Jensen, 1969). IQ tests have gone to court, provoked legislation, and taken thrashings from the popular media (Brown, Reynolds, & Whitaker, 1999; Reynolds, 2000a). In New York, the conflict has culminated in laws known as truth-in-testing legislation, which some clinicians say interferes with professional practice. In California, a ban

This chapter is based substantively on a chapter that appears in the prior edition of this text by Reynolds and Ramsay (2003).

was placed on the use of IQ tests for identification and placement of African American students.

In statistics, *bias* refers to systematic error in the estimation of a value. A biased test is one that systematically overestimates or underestimates the value of the variable it is intended to assess. If this bias occurs as a function of a nominal cultural variable, such as ethnicity or gender, cultural test bias is said to be present. On the Wechsler series of intelligence tests, for example, the difference in mean scores for Black and White Americans hovers around 15 points. If this figure represents a true difference between the two groups, the tests are not biased. If, however, the difference is due to systematic underestimation of the intelligence of Black Americans or overestimation of the intelligence of White Americans, the tests are said to be culturally biased.

Many researchers have investigated possible bias in intelligence tests, with inconsistent results. The question of test bias remained chiefly within the purview of scientists until the 1970s. Since then it has become a major social issue, touching off heated public debate (e.g., Brooks, 1997; Fine, 1975). Many professionals and professional associations have taken strong stands on the question. Van de Vijver and Tanzer (2004) presented a taxonomy of three kinds of bias:

1. *Construct bias* [e.g., overlap in definitions of the construct across cultures, "differential appropriateness of behaviors associated with the construct in different cultures" (p. 124), and poor sampling of relevant behaviors associated with the construct]
2. *Method bias* [i.e., bias pertaining to the sample (e.g., samples are not matched in terms of all relevant aspects, which is nearly impossible to achieve), instrument (e.g., differential familiarity with the items), or administration (e.g., ambiguous directions, tester/interviewer/observer effects)]
3. *Item bias* due to "poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, or influence of culture specifics such as nuisance factors or connotations associated with the item wording" (p. 127).

A number of strategies are available to address bias in cross-cultural assessment (van de Vijver & Tanzer, 2004).

## MINORITY OBJECTIONS TO TESTS AND TESTING

Since 1968, the Association of Black Psychologists (ABP) has called for a moratorium on the administration of psychological and educational tests with minority examinees (Samuda, 1975; Williams, Dotson, Dow, & Williams, 1980). The ABP brought this call to other professional associations in psychology and education. The American Psychological Association (APA) responded by requesting that its Board of Scientific Affairs establish a committee to study the use of these tests with disadvantaged students. (See the committee's report, Cleary, Humphreys, Kendrick, & Wesman, 1975.)

The ABP published this policy statement in 1969 (cited in Williams et al., 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) potentiate inferior education; (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 children. (pp. 265–266)

Subsequently, other professional associations issued policy statements on testing. Williams et al. (1980) and Reynolds, Lowe, and Saenz (1999) cited the National Association for the Advancement of Colored People (NAACP), the National Education Association, the National Association of Elementary School Principals, and the American Personnel and Guidance Association, among others, as organizations releasing such statements.

The ABP, perhaps motivated by action and encouragement on the part of the NAACP, adopted a more detailed resolution in 1974. The resolution described, in part, these goals of the ABP: (a) a halt to the standardized testing of Black people until culture-specific tests are made available, (b) a national policy of testing by competent assessors of an examinee's own ethnicity at his or her mandate, (c) removal of standardized test results from the records of Black students and employees, and (d) a return to regular programs of Black students inappropriately diagnosed and placed in special education classes (Williams et al., 1980). This statement presupposes that flaws in standardized tests are responsible for the unequal test results of Black examinees and, with them, any detrimental consequences of those results. Concerns continue despite the 2004 reauthorization of the Individuals with Disabilities Education Act, which indicates alternative methods can be used (e.g., Response to Intervention [RTI] to assess learning disabilities eliminating reliance upon an intelligence/achievement

discrepancy formula. A study of African American psychology professionals indicated that concerns remain as to whether RTI will reduce the disproportionately high number of African Americans students in special education (Graves & Mitchell, 2011).

## ORIGINS OF THE TEST BIAS CONTROVERSY

Challenges of test bias have emerged given the emphasis placed on American societal values and beliefs, the nature of tests and testing, and conflicting views regarding definition.

### Social Values and Beliefs

The present-day conflict over bias in standardized tests is motivated largely by public concerns. The impetus, it may be argued, lies with beliefs fundamental to democracy in the United States. Most Americans, at least those of majority ethnicity, view the United States as a land of opportunity. Historically, this has meant that equal opportunity is extended to every person.

We want to believe that any child can grow up to be president. Concomitantly, we believe that everyone is created equal, that all people harbor the potential for success and achievement. This equality of opportunity seems most reasonable if everyone is equally able to take advantage of it. Concerns have arisen given debates among scholars as to whether intelligence is a fixed trait (i.e., corresponding test scores are stable over time) or whether intelligence is malleable (Ramsden et al., 2011; Suzuki & Aronson, 2005).

Parents and educational professionals have corresponding beliefs: The children we serve have an immense potential for success and achievement; the great effort we devote to teaching or raising children is effort well spent; my own child is intelligent and capable. The result is a resistance to labeling and alternative placement, which are thought to discount students' ability and diminish their opportunity. This terrain may be a bit more complex for clinicians, because certain diagnoses have consequences desired by clients. A disability diagnosis, for example, allows people to receive compensation or special services, and insurance companies require certain serious conditions for coverage.

### Character of Tests and Testing

The nature of psychological characteristics and their measurement is partly responsible for long-standing concern over test bias (Reynolds & Brown, 1984a). Psychological characteristics are internal, so scientists cannot observe or measure them directly but must infer them from a person's external behavior. By extension, clinicians must contend with the same limitation.

According to MacCorquodale and Meehl (1948), a psychological process is an *intervening variable* if it is treated only as a component of a system and has no properties beyond the ones that operationally define it. It is a *hypothetical construct* if it is thought to exist and to have properties beyond its defining ones. In biology, a *gene* is an example of a hypothetical construct. The gene has properties beyond its use to describe the transmission of traits from one generation to the next. Both intelligence and personality have the status of hypothetical constructs. The nature of psychological processes and other unseen hypothetical constructs are often subjects of persistent debate. (See Ramsay, 1998b, for one approach.) Intelligence, a highly complex, multifaceted psychological process, has given rise to disputes that are especially difficult to resolve (Reynolds, Willson, & Ramsey, 1999). Test development procedures (Ramsay & Reynolds, 2000a) are essentially the same for all standardized tests. Initially, the author of a test develops or collects a large pool of items thought to measure the characteristic of interest. Theory and practical usefulness are standards commonly used to select an item pool. The selection process is a rational one. That is, it depends on reason and judgment. A rigorous means to carry out the item selection process at this stage simply does not exist. At this stage, then, test authors have no generally accepted evidence that they have selected appropriate items.

A common second step is to discard items of suspect quality, again on rational grounds, to reduce the pool to a manageable size. Next, the test's author or publisher administers the items to a group of examinees called a *tryout sample*. Statistical procedures then help to identify items that seem to be measuring an unintended characteristic or more than one characteristic. The author or publisher discards or modifies these items.

Finally, examiners administer the remaining items to a large, diverse group of people called a standardization sample or *norming sample*. This sample should reflect every important characteristic of the population that will take the final version of the test. Statisticians compile the scores of the norming sample into an array called a *norming distribution*. In order to address concerns regarding racial and ethnic group representation in the norming sample, some test developers engage in racial and ethnic group oversampling (i.e., including larger numbers of

individuals from different racial and ethnic groups above and beyond their proportional representation in the overall population). Supplemental norms may then be created for a particular racial/ethnic group. Tests such as the Wechsler scales often incorporate this oversampling procedure (cited in Suzuki, Kugler, & Aguiar, 2005).

Eventually, clients or other examinees take the test in its final form. The scores they obtain, known as *raw scores*, do not yet have any interpretable meaning. A clinician compares these scores with the norming distribution. The comparison is a mathematical process that results in new, *standard scores* for the examinees. Clinicians can interpret these scores, whereas interpretation of the original, raw scores would be difficult and impractical in the absence of a comparable norm group (Reynolds, Lowe, et al., 1999).

Standard scores are relative. They have no meaning in themselves but derive their meaning from certain properties—typically the mean and standard deviation (*SD)* of the norming distribution. The norming distributions of many ability tests, for example, have a mean score of 100 and a standard deviation of 15. A client might obtain a standard score of 127. This score would be well above average, because 127 is almost 2 SDs of 15 above the mean of 100. Another client might obtain a standard score of 96. This score would be a little below average, because 96 is about one third of a SD below a mean of 100.

Here the reason why raw scores have no meaning gains a little clarity. A raw score of, say, 34 is high if the mean is 30 but low if the mean is 50. It is very high if the mean is 30 and the SD is 2 but less high if the mean is again 30 and the SD is 15. Thus, a clinician cannot know how high or low a score is without knowing certain properties of the norming distribution. The standard score is the one that has been compared with this distribution, so that it reflects those properties. (See Ramsay & Reynolds, 2000a, for a systematic description of test development.)

Charges of bias frequently spring from low proportions of minorities in the norming sample of a test and correspondingly small influence on test results. Many norming samples include only a few minority participants, eliciting suspicion that the tests produce inaccurate scores—misleadingly low ones in the case of ability tests—for minority examinees. Whether this is so is an important question that calls for scientific study (Reynolds, Lowe et al., 1999).

Test development is a complex and elaborate process (Ramsay & Reynolds, 2000a). The public, the media, Congress, and even the intelligentsia find it difficult to understand. Clinicians, and psychologists outside the measurement field, commonly have little knowledge of the issues surrounding this process. Its abstruseness, as much as its relative nature, probably contributes to the amount of conflict over test bias. Physical and biological measurements such as height, weight, and even risk of heart disease elicit little controversy, although they vary from one ethnic group to another. As explained by Reynolds, Lowe et al. (1999), this is true in part because such measurements are absolute, in part because they can be obtained and verified in direct and relatively simple ways, and in part because they are free from the distinctive social implications and consequences of standardized test scores. Reynolds et al. correctly suggested that test bias is a special case of the uncertainty that accompanies all measurement in science. Ramsay (2000) and Ramsay and Reynolds (2000b) presented a brief treatment of this uncertainty incorporating Heisenberg's model.

## Divergent Ideas of Bias

Besides the character of psychological processes and their measurement, differing understandings held by various segments of the population also add to the test bias controversy. Researchers and laypeople view bias differently. Clinicians and other professionals bring additional divergent views. Many lawyers see bias as illegal, discriminatory practice on the part of organizations or individuals (Reynolds, 2000a; Reynolds & Brown, 1984a).

To the public at large, bias sometimes conjures up notions of prejudicial attitudes. A person seen as prejudiced may be told, "You're biased against Hispanics." For other laypersons, bias is more generally a characteristic slant in another person's thinking, a lack of objectivity brought about by the person's life circumstances. A sales clerk may say, "I think sales clerks should be better paid." "Yes, but you're biased," a listener may retort. These views differ from statistical and research definitions for bias as for other terms, such as *significant, association,* and *confounded.* The highly specific research definitions of such terms are unfamiliar to almost everyone. As a result, uninitiated readers often misinterpret research reports.

Both in research reports and in public discourse, the scientific and popular meanings of bias are often conflated, as if even the writer or speaker had a tenuous grip on the distinction. Reynolds, Lowe et al. (1999) have suggested that the topic would be less controversial if research reports addressing test bias as a scientific question relied on the scientific meaning alone.

## EFFECTS AND IMPLICATIONS OF THE TEST BIAS CONTROVERSY

The dispute over test bias has given impetus to an increasingly sophisticated corpus of research. In most venues, tests of reasonably high statistical quality appear to be largely unbiased. For neuropsychological tests, results are not definitive but so far they appear to indicate little bias. Studies examining psychophysiological approaches to intelligence have yielded results that may further elucidate the relationship between culture and cognitive functioning. Verney, Granholm, Marshall, Malcarne, and Saccuzzo (2005) found that measures of information processing efficiency were related to Caucasian American students' performance but not to a comparable sample of Mexican American students, suggesting differential validity in prediction. Both sides of the debate have disregarded most of these findings and have emphasized, instead, a mean difference between ethnic groups (Reynolds, 2000b).

In addition, publishers have released nonverbal tests that have been identified as culture-reduced measures of ability; practitioners interpret scores so as to minimize the influence of putative bias; and, finally, publishers revise tests directly, to expunge group differences. For minority group members, these revisions may have an undesirable long-range effect: to prevent the study and thereby the remediation of any bias that might otherwise be found. In addition, all tests involve some form of language and communication (Mpofu & Ortiz, 2009). Methods for detecting bias on nonverbal measures are the same as those for tests with verbal content. Information regarding bias studies on several nonverbal measures including the Comprehensive Test of Nonverbal Intelligence (CTONI; Hammill, Pearson, & Wiederholt, 1997), Leiter International Performance Scale-R (LIPS-R; Roid & Miller, 1997), and Universal Nonverbal Intelligence Test (UNIT: Bracken & McCallum, 1998) are evaluated in seminal texts (Maller, 2003). Results indicate that differences in performance by racial and ethnic groups are reduced on nonverbal measures.

The implications of these various effects differ depending on whether the bias explanation is correct or incorrect, assuming it is accepted. An incorrect bias explanation, if accepted, would lead to modified tests that would not reflect important, correct information and, moreover, would present the incorrect information that unequally performing groups had performed equally. Researchers, unaware or unmindful of such inequalities, would neglect research into the causes of these inequalities. Economic and social deprivation would come to appear less harmful and therefore more justifiable. Social programs, no longer seen as necessary to improve minority students' scores, might be discontinued, with serious consequences.

A correct bias explanation, if accepted, would leave professionals and minority group members in a relatively better position. We would have copious research correctly indicating that bias was present in standardized test scores. Surprisingly, however, the limitations of having these data might outweigh the benefits. Test bias would be a correct conclusion reached incorrectly.

Findings of bias rely primarily on mean differences between groups. These differences would consist partly of bias and partly of other constituents, which would project them upward or downward, perhaps depending on the particular groups involved. Thus, we would be accurate in concluding that bias was present but inaccurate as to the amount of bias and, possibly, its direction: that is, which of two groups it favored. Any modifications made would do too little or too much, creating new bias in the opposite direction.

The presence of bias should allow for additional explanations. For example, bias and *Steelean effects* (Steele & Aronson, 1995, 2004), in which fear of confirming a stereotype impedes minorities' performance, might both affect test results. Research indicates that stereotype threat may be viewed as a source of measurement bias (Wicherts, Dolan, & Hessen, 2005). Such additional possibilities, which now receive little attention, would receive even less. Economic and social deprivation, serious problems apart from testing issues, would again appear less harmful and therefore more justifiable. Efforts to improve people's scores through social programs would be difficult to defend, because this work presupposes that factors other than test bias are the causes of score differences. Thus, Americans' belief in human potential would be vindicated, but perhaps at considerable cost to minority individuals.

## POSSIBLE SOURCES OF BIAS

Minority and other psychologists have expressed numerous concerns over the use of psychological and educational tests with minorities. These concerns are potentially legitimate and substantive but are often asserted as true in the absence of scientific evidence. Reynolds, Lowe et al. (1999) have divided the most frequent of the problems cited into seven categories, described briefly here. Two categories, inequitable social consequences and qualitatively distinct aptitude and personality, receive

more extensive treatments in the "Test Bias and Social Issues" section.

1. *Inappropriate content*. Tests are geared to majority experiences and values or are scored arbitrarily according to majority values. Correct responses or solution methods depend on material that is unfamiliar to minority individuals.
2. *Inappropriate standardization samples*. Minorities' representation in norming samples is proportionate but insufficient to allow them any influence over test development.
3. *Examiners' and language bias*. White examiners who speak standard English intimidate minority examinees and communicate inaccurately with them, spuriously lowering their test scores.
4. *Inequitable social consequences*. Ethnic minority individuals, already disadvantaged because of stereotyping and past discrimination, are denied employment or relegated to dead-end educational tracks. Labeling effects are another example of invalidity of this type.
5. *Measurement of different constructs*. Tests largely based on majority culture are measuring different characteristics altogether for members of minority groups, rendering them invalid for these groups.
6. *Differential predictive validity*. Standardized tests accurately predict many outcomes for majority group members, but they do not predict any relevant behavior for their minority counterparts. In addition, the criteria that tests are designed to predict, such as achievement in White, middle-class schools, may themselves be biased against minority examinees.
7. *Qualitatively distinct aptitude and personality*. This position seems to suggest that minority and majority ethnic groups possess characteristics of different *types*, so that test development must begin with different definitions for majority and minority groups.

Researchers have investigated these concerns, although few results are available for labeling effects or for long-term social consequences of testing. As noted by Reynolds, Lowe et al. (1999), both of these problems are relevant to testing in general rather than to ethnic issues alone. In addition, individuals as well as groups can experience labeling and other social consequences of testing. Researchers should investigate these outcomes with diverse samples and numerous statistical techniques. Finally, Reynolds, Lowe et al. suggest that tracking and special education should be treated as problems with education rather than assessment.

## WHAT TEST BIAS IS AND IS NOT

Scientists and clinicians should distinguish bias from *unfairness* and from *offensiveness*. Thorndike (1971) wrote, "The presence (or absence) of differences in mean score between groups, or of differences in variability, tells us nothing directly about fairness" (p. 64). In fact, the concepts of test bias and unfairness are distinct in themselves. A test may have very little bias, but a clinician could still use it unfairly to minority examinees' disadvantage. Conversely, a test may be biased, but clinicians need not—and must not—use it to unfairly penalize minorities or others whose scores may be affected. Little is gained by anyone when concepts are conflated or when, in any other respect, professionals operate from a base of misinformation.

Jensen (1980) was the author who first argued cogently that fairness and bias are separable concepts. As noted by Brown et al. (1999), fairness is a moral, philosophical, or legal issue on which reasonable people can legitimately disagree. By contrast, bias is an empirical property of a test, as used with two or more specified groups. Thus, bias is a statistically estimated quantity rather than a principle established through debate and opinion.

A second distinction is that between test bias and item *offensiveness*. In the development of many tests, a minority review panel examines each item for content that may be offensive to one or more groups. Professionals and laypersons alike often view these examinations as tests of bias. Such *expert reviews* have been part of the development of many prominent ability tests, including the Kaufman Assessment Battery for Children (K-ABC), the Wechsler Preschool and Primary Scale of Intelligence–Revised (WPPSI-R), and the Peabody Picture Vocabulary Test–Revised (PPVT-R). The development of personality and behavior tests also incorporates such reviews (e.g., Reynolds, 2001; Reynolds & Kamphaus, 1992). Prominent authors such as Anastasi (1988), Kaufman (1979), and Sandoval and Mille (1979) support this method as a way to enhance rapport with the public.

In a well-known case titled *PASE v. Hannon* (Reschly, 2000), a federal judge applied this method rather quaintly, examining items from the Wechsler Intelligence Scales for Children (WISC) and the Binet intelligence scales to personally determine which items were biased (Elliot, 1987). Here an authority figure showed startling naiveté and greatly exceeded his expertise—a telling comment on modern hierarchies of influence. Similarly, a high-ranking representative of the Texas Education Agency argued in a televised interview (October 14, 1997, KEYE 42, Austin, TX) that the Texas Assessment of Academic

Skills (TAAS), controversial among researchers, could not be biased against ethnic minorities because minority reviewers inspected the items for biased content.

Several researchers have reported that such expert reviewers perform at or below chance level, indicating that they are unable to identify biased items (Jensen, 1976; Sandoval & Mille, 1979; reviews by Camilli & Shepard, 1994; Reynolds, 1995, 1998a; Reynolds, Lowe et al., 1999). Since initial research by McGurk (1951), studies have provided little evidence that anyone can estimate, by personal inspection, how differently a test item may function for different groups of people.

Sandoval and Mille (1979) had university students from Spanish, history, and education classes identify items from the WISC-R that would be more difficult for a minority child than for a White child, along with items that would be equally difficult for both groups. Participants included Black, White, and Mexican American students. Each student judged 45 items, of which 15 were most difficult for Blacks, 15 were most difficult for Mexican Americans, and 15 were most nearly equal in difficulty for minority children, in comparison with White children.

The participants read each question and identified it as easier, more difficult, or equally difficult for minority versus White children. Results indicated that the participants could not make these distinctions to a statistically significant degree and that minority and nonminority participants did not differ in their performance or in the types of misidentifications they made. Sandoval and Mille (1979) used only extreme items, so the analysis would have produced statistically significant results for even a relatively small degree of accuracy in judgment.

For researchers, test bias is a deviation from examinees' real level of performance. Bias goes by many names and has many characteristics, but it always involves scores that are too low or too high to accurately represent or predict some examinee's skills, abilities, or traits. To show bias, then—to greatly simplify the issue—requires estimates of scores. Reviewers have no way of producing such an estimate.

Despite these issues regarding fairness and expert reviews, testing companies continue to employ these methods as part of the development process. For example, the Educational Testing Service (2002) provides *Standards for Quality and Fairness* and *International Principles for Fairness Review of Assessments*. These documents emphasize the importance of: treating people with respect; minimizing the effects of construct-irrelevant knowledge or skills; avoiding material that is unnecessarily controversial, inflammatory, offensive, or upsetting; using appropriate terminology; avoiding stereotypes; and representing diversity (Zieky, 2006).

While these procedures can suggest items that may be offensive, statistical techniques are necessary to determine test bias. Thus, additional procedures for fairness reviews include a focus on how to resolve disputes among reviewers, attention to test design, diverse input, provision of accommodations, differential item functioning, validation of the review process, and attention to how the test is used (Zieky, 2006).

### Culture Fairness, Culture Loading, and Culture Bias

A third pair of distinct concepts is cultural *loading* and cultural *bias*, the former often associated with the concept of culture fairness. Cultural loading is the degree to which a test or item is specific to a particular culture. A test with greater cultural loading has greater potential bias when administered to people of diverse cultures. Nevertheless, a test can be culturally loaded without being culturally biased.

An example of a culture-loaded item might be "Who was Eleanor Roosevelt?" This question may be appropriate for students who have attended U.S. schools since first grade with curriculum highlighting her importance as a historical figure in America. The cultural specificity of the question would be too great, however, to permit its use with European and certainly Asian elementary school students, except perhaps as a test of knowledge of U.S. history. Nearly all standardized tests have some degree of cultural specificity. Cultural loadings fall on a continuum, with some tests linked to a culture as defined very generally and liberally and others to a culture as defined very narrowly.

Cultural loading, by itself, does not render tests biased or offensive. Rather, it creates a potential for either problem, which must then be assessed through research. Ramsay (2000; Ramsay & Reynolds, 2000b) suggested that some characteristics might be viewed as desirable or undesirable in themselves but others as desirable or undesirable only to the degree that they influence other characteristics. Test bias against Cuban Americans would itself be an undesirable characteristic. A subtler situation occurs if a test is both culturally loaded and culturally biased. If the test's cultural loading is a cause of its bias, the cultural loading is then *indirectly* undesirable and should be corrected. Alternatively, studies may show that the test is culturally loaded but unbiased. If so, indirect undesirability due to an association with bias can be ruled out.

Some authors (e.g., Cattell, 1979) have attempted to develop culture-fair intelligence tests. These tests, however,

are characteristically poor measures from a statistical standpoint (Anastasi, 1988; Ebel, 1979). In one study, Hartlage, Lucas, and Godwin (1976) compared Raven's Progressive Matrices (RPM), thought to be culture fair, with the WISC, thought to be culture loaded. The researchers assessed these tests' predictiveness of reading, spelling, and arithmetic measures with a group of disadvantaged, rural children of low socioeconomic status (SES). WISC scores consistently correlated higher than RPM scores with the measures examined.

The problem may be that intelligence is defined as adaptive or beneficial behavior within a particular culture. Therefore, a test free from cultural influence would tend to be free from the influence of intelligence—and to be a poor predictor of intelligence in any culture. As Reynolds, Lowe et al. (1999) observed, if a test is developed in one culture, its appropriateness to other cultures is a matter for scientific verification. Test scores should not be given the same interpretations for different cultures without evidence that those interpretations would be sound.

### Test Bias and Social Issues

Authors have introduced numerous concerns regarding tests administered to ethnic minorities (Brown et al., 1999). Many of these concerns, however legitimate and substantive, have little connection with the scientific estimation of test bias. According to some authors, the unequal results of standardized tests produce inequitable social consequences. Low test scores relegate minority group members, already at an educational and vocational disadvantage because of past discrimination and low expectations of their ability, to educational tracks that lead to mediocrity and low achievement (Chipman, Marshall, & Scott, 1991; Payne & Payne, 1991; see also "Possible Sources of Bias" section).

Other concerns are more general. Proponents of tests, it is argued, fail to offer remedies for racial or ethnic differences (Scarr, 1981), to confront societal concerns over racial discrimination when addressing test bias (Gould, 1995, 1996), to respect research by cultural linguists and anthropologists (Figueroa, 1991; Helms, 1992), to address inadequate special education programs (Reschly, 1997), and to include sufficient numbers of African Americans in norming samples (Dent, 1996). Furthermore, test proponents use massive empirical data to conceal historic prejudice and racism (Richardson, 1995). Some of these practices may be deplorable, but they do not constitute test bias. A removal of group differences from scores cannot combat them effectively and may even remove some evidence of their existence or influence.

Gould (1995, 1996) has acknowledged that tests are not statistically biased and do not show differential predictive validity. He has argued, however, that defining cultural bias statistically is confusing: The public is concerned not with statistical bias but with whether Black–White IQ differences occur because society treats Black people unfairly. That is, the public considers tests biased if they record biases originating elsewhere in society (Gould, 1995). Researchers consider them biased only if they introduce additional error because of flaws in their design or properties. Gould (1995, 1996) argued that society's concern cannot be addressed by demonstrations that tests are statistically unbiased. It can, of course, be addressed empirically.

Another social concern, noted briefly earlier, is that majority and minority examinees may have qualitatively different aptitudes and personality traits, so that traits and abilities must be conceptualized differently for different groups. If this is not done, a test may produce lower results for one group because it is conceptualized most appropriately for another group. This concern is complex from the standpoint of construct validity and may take various practical forms.

In one possible scenario, two ethnic groups can have different patterns of abilities, but the sums of their abilities can be about equal. Group A may have higher verbal fluency, vocabulary, and usage but lower syntax, sentence analysis, and flow of logic than Group B. A verbal ability test measuring only the first three abilities would incorrectly represent Group B as having lower verbal ability. This concern is one of construct validity.

Alternatively, a verbal fluency test may be used to represent the two groups' verbal ability. The test accurately represents Group B as having lower verbal fluency but is used inappropriately to suggest that this group has lower verbal ability per se. Such a characterization is not only incorrect; it is unfair to group members and has detrimental consequences for them that cannot be condoned. Construct invalidity is difficult to argue here, however, because this concern is one of test use.

## RELATED QUESTIONS

The next sections clarify what can be inferred from test score differences and the application of statistical methods to investigate test bias.

### Test Bias and Etiology

The etiology of a condition is distinct from the question of test bias. (For a review, see Reynolds & Kaiser, 1992.) In fact, the need to research etiology emerges only after evidence that a score difference is a real one, not an artifact of bias. Authors have sometimes inferred that score differences themselves indicate genetic differences, implying that one or more groups are genetically inferior. This inference is scientifically no more defensible—and ethically much less so—than the notion that score differences demonstrate test bias.

Jensen (1969) has long argued that mental tests measure, to some extent, the intellectual factor *g*, found in behavioral genetics studies to have a large genetic component. In Jensen's view, group differences in mental test scores may reflect largely genetic differences in *g*. Nonetheless, Jensen made many qualifications to these arguments and to the differences themselves. He also posited that other factors make considerable, though lesser, contributions to intellectual development (Reynolds, Lowe et al., 1999). Jensen's theory, if correct, may explain certain intergroup phenomena, such as differential Black and White performance on digit span measures (Ramsay & Reynolds, 1995).

### Test Bias Involving Groups and Individuals

Bias may influence the scores of individuals as well as groups on personality and ability tests. Therefore, researchers can and should investigate both of these possible sources of bias. An overarching statistical method called the general linear model permits this approach by allowing both *group* and *individual* to be analyzed as independent variables. In addition, item characteristics, motivation, and other nonintellectual variables (Reynolds, Lowe et al. 1999; Sternberg, 1980; Wechsler, 1975) admit of analysis through recoding, categorization, and similar expedients.

## EXPLAINING GROUP DIFFERENCES

Among researchers, the issue of cultural bias stems largely from well-documented findings, now seen in more than 100 years of research, that members of different ethnic groups have different levels and patterns of performance on many prominent cognitive ability tests. Intelligence batteries have generated some of the most influential and provocative of these findings (Elliot, 1987; Gutkin & Reynolds, 1981; Reynolds, Chastain, Kaufman, & McLean, 1987; Spitz, 1986). In many countries worldwide, people of different ethnic and racial groups, genders, socioeconomic levels, and other demographic groups obtain systematically different intellectual test results. Black–White IQ differences in the United States have undergone extensive investigation for more than 50 years. Jensen (1980), Shuey (1966), Tyler (1965), and Willerman (1979) have reviewed the greater part of this research. The findings occasionally differ somewhat from one age group to another, but they have not changed substantially in the past century. Scholars often refer to the racial and ethnic group hierarchy of intelligence that has remained in the same consistent order for decades. Overall estimates based on a mean of 100 and *SD* of 15 are often cited in this way: Whites 100, Black/African Americans 85, Hispanics midway between Whites and Blacks; Asians and Jews above 100 ("Mainstream Science on Intelligence," 1994). American Indians score at approximately 90 (McShane, 1980).

On average, Blacks differ from Whites by about 1.0 *SD*, with White groups obtaining the higher scores. The differences have been relatively consistent in size for some time and under several methods of investigation. An exception is a reduction of the Black–White IQ difference on the intelligence portion of the K-ABC to about .5 *SD*s, although this result is controversial and poorly understood. (See Kamphaus & Reynolds, 1987, for a discussion.) In addition, such findings are consistent only for African Americans. Other highly diverse findings appear for native African and other Black populations (Jensen, 1980).

Researchers have taken into account a number of demographic variables, most notably SES. The size of the mean Black–White difference in the United States then diminishes to .5 to .7 *SD*s (Jensen, 1980; Kaufman, 1973; Kaufman & Kaufman, 1973; Reynolds & Gutkin, 1981) but is robust in its appearance. It should be noted that mean score differences between Black and White Americans have lessened over the years. For example, IQ differences between Black and White 12-year-olds have dropped 5.5 points to 9.5 over the past three decades (Nisbett, 2009).

While group differences have often received attention in the literature, differences in general ability areas, such as verbal and spatial abilities, are also noted within particular racial and ethnic groups. For example, Suzuki, et al. (2005) conducted a preliminary analysis of Wechsler studies including American Indian samples between 1986 and 2003. A total of 63 studies included samples from the Navajo, Papago, Ojibwa, Inuit, and Eskimo communities. All studies revealed higher performance on nonverbal spatial reasoning tasks (e.g., Object Assembly and Block Design) in comparison to verbal subtests (e.g., Information and Vocabulary). The standard score difference between Verbal IQ and Performance IQ was

approximately 17 points (SD 8.92). Explanation of these findings focused on the Verbal IQ being lower due to linguistic and cultural factors, thus leading authors to suggest that the Performance IQ may be more indicative of intellectual potential in American Indian communities. Hagie, Gallipo, and Svien (2003) examined American Indian students' patterns of performance across items on the Bayley Scales of Infant Development (BSID) and the WISC-III. The authors reported based on their analysis that "[i]ssues of poverty, remoteness, access to resources, and health care need to be considered before sweeping conclusions can be made about performance on nationally normed, standardized instruments" (p. 15). In addition, they concluded that these traditional measures may yield "distorted and inaccurate results due to cultural biases of test items and environmental concerns" (p. 24).

Asian groups, although less thoroughly researched than Black groups, have consistently performed as well as or better than Whites (Pintner, 1931; Tyler, 1965; Willerman, 1979). Asian Americans obtain average mean ability scores (Flynn, 1991; Lynn, 1995; Neisser et al., 1996; Reynolds, Willson, et al., 1999). It is important to note that most of the published studies in the past decade have focused on non–U.S. international Asian samples (Okazaki & Sue, 2000). The demand for intelligence tests like the Wechsler scales in Asia has led to the exporting of measures that are then normed and restandardized. For example, the WAIS has been translated and standardized in China, Hong Kong, India, Japan, Korea, Taiwan, Thailand, and Vietnam (Cheung, Leong, & Ben-Porath, 2003).

Matching is an important consideration in studies of ethnic differences. Any difference between groups may be due to neither test bias nor ethnicity but to SES, nutrition, home environment, and other variables that may be associated with test performance. Matching on these variables controls for their associations.

A limitation to matching is that it results in regression toward the mean. Black respondents with high self-esteem, for example, may be selected from a population with low self-esteem. When examined later, these respondents will test with lower self-esteem, having regressed to the lower mean of their own population. Their extreme scores—high in this case—were due to chance.

Clinicians and research consumers should also be aware that the similarities between ethnic groups are much greater than the differences. This principle holds for intelligence, personality, and most other characteristics, both psychological and physiological. From another perspective, the variation among members of any one ethnic group greatly exceeds the differences between groups. The large similarities among groups appear repeatedly in analyses as large, statistically significant constants and great overlap between different groups' ranges of scores.

Some authors (e.g., Schoenfeld, 1974) have disputed whether racial differences in intelligence are real or even researchable. Nevertheless, the findings are highly reliable from study to study, even when study participants identify their own race. Thus, the existence of these differences has gained wide acceptance. The differences are real and undoubtedly complex. The tasks remaining are to describe them thoroughly (Reynolds, Lowe et al., 1999) and, more difficult, to explain them in a causal sense (Ramsay, 1998a, 2000). Both the lower scores of some groups and the higher scores of others must be explained, and not necessarily in the same way.

Over time, exclusively genetic and environmental explanations have lost so much of their credibility that they can hardly be called current. Most researchers who posit that score differences are real now favor an interactionist perspective. This development reflects a similar shift in psychology and social science as a whole. However, this relatively recent consensus masks the subtle persistence of an earlier assumption that test score differences must have either a genetic or an environmental basis. The relative contributions of genes and environment still provoke debate, with some authors seemingly intent on establishing a predominantly genetic or a predominantly environmental basis. The interactionist perspective shifts the focus of debate from *how much* to *how* genetic and environmental factors contribute to a characteristic. In practice, not all scientists have made this shift. In 2005, Rushton and Jensen published a monograph focusing on the past 30 years of research on race differences in cognitive ability. The culture-only (0% genetic, 100% environmental) and hereditarian (50% genetic 50% environmental) perspectives were examined based on a variety of sources of evidence, including: worldwide distribution of test scores, *g* factor of mental ability, brain size and cognitive ability, transracial adoption studies, human origins research, and hypothesized environmental variables. Rushton and Jensen concluded that their extensive findings support a hereditarian explanation for race differences. A number of scholars, however, debated their findings and the interpretation of the data from a number of studies.

## CULTURAL TEST BIAS AS AN EXPLANATION

The bias explanation of score differences has led to the cultural test bias hypothesis (CTBH; Brown et al., 1999;

Reynolds, 1982a, 1982b; Reynolds & Brown, 1984b). According to the CTBH, differences in mean performance for members of different ethnic groups do not reflect real differences among groups but are artifacts of tests or of the measurement process. This approach holds that ability tests contain systematic error occurring as a function of group membership or other nominal variables that should be irrelevant. That is, people who should obtain equal scores obtain unequal ones because of their ethnicities, genders, socioeconomic levels, and the like.

For SES, Eells, Davis, Havighurst, Herrick, and Tyler (1951) summarized the logic of the CTBH in this way: If (a) children of different SES levels have experiences of different kinds and with different types of material, and if (b) intelligence tests contain a disproportionate amount of material drawn from cultural experiences most familiar to high-SES children, then (c) high-SES children should have higher IQ scores than low-SES children. As Eells et al. observed, this argument tends to imply that IQ differences are artifacts that depend on item content and "do not reflect accurately any important underlying ability" (p. 4) in the individual. Sattler (2008) noted that "poverty in and of itself is not necessary nor sufficient to produce intellectual deficits," although children growing up in this context may be exposed to "low level parental education, poor nutrition and health care, substandard housing, family disorganization, inconsistent discipline, diminished sense of personal worth, low expectations, frustrated aspirations, physical violence in their neighborhoods, and other environmental pressures" (pp. 137–138).

Since the 1960s, the CTBH explanation has stimulated numerous studies, which in turn have largely refuted the explanation. Lengthy reviews are available (e.g., Jensen, 1980; Reynolds, 1995, 1998a; Reynolds & Brown, 1984b). This literature suggests that tests whose development, standardization, and reliability are sound and well documented are not biased against native-born American racial or ethnic minorities. Studies do occasionally indicate bias, but it is usually small, and most often it favors minorities.

Results cited to support content bias indicate that item biases account for $< 1\%$ to about 5% of variation in test scores. In addition, it is usually counterbalanced across groups. That is, when bias against an ethnic group occurs, comparable bias favoring that group occurs also and cancels it out. When apparent bias is counterbalanced, it may be random rather than systematic and therefore not bias after all. Item or subtest refinements, as well, frequently reduce and counterbalance bias that is present.

No one explanation is likely to account for test score differences in their entirety. A contemporary approach to statistics, in which effects of zero are rare or even nonexistent, suggests that tests, test settings, and nontest factors may all contribute to group differences. (See also Bouchard & Segal, 1985; Flynn, 1991; Loehlin, Lindzey, & Spuhler, 1975.)

Some authors, most notably Mercer (1979; see also Helms, 1992; Lonner, 1985), have reframed the test bias hypothesis over time. Mercer argued that the lower scores of ethnic minorities on aptitude tests can be traced to the Anglocentrism, or adherence to White, middle-class value systems, of these tests. Mercer's assessment system, the System of Multicultural Pluralistic Assessment (SOMPA), effectively equated ethnic minorities' intelligence scores by applying complex demographic corrections. The SOMPA was popular for several years. It is used less commonly today because of its conceptual and statistical limitations (Reynolds, Lowe et al., 1999). Gopaul-McNicol and Armour-Thomas (2002) proposed a biocultural assessment system incorporating psychometric assessment, psychometric potential assessment [i.e., "value added information about nascent potentials not yet fully developed or competencies not likely to be determined under standardized testing conditions" (p. 38)], ecological assessment (direct observation in the relevant contexts of the individual), and other intelligences assessment (cognitive strengths beyond the IQ test).

In addition, the Gf-Gc Cross-Battery Assessment Model (XBA; Flanagan, Ortiz, & Alfonso, 2007) takes into consideration a wider range of cognitive abilities enabling the evaluator to select from a range of potential tests, addressing broad and narrow ability areas, rather than relying on one battery of subtests (McGrew & Flanagan, 1998). As part of this model, the authors developed the Culture-Language Test Classifications (C-LTC; McGrew & Flanagan, 1998). The C-LTC is based on the degree of cultural loading (i.e., cultural specificity) and linguistic demand of various measures. The classification is based on examination of empirical data available as well as expert consensus procedures when data are not available. The Culture-Language Interpretive Matrix (C-LIM) is derived from this classification system and is represented by a matrix to assist the evaluator in test selection and interpretation (Ortiz & Ochoa, 2005). The model takes into consideration issues of acculturation and language proficiency.

## HARRINGTON'S CONCLUSIONS

Unlike such authors as Mercer (1979) and Helms (1992), Harrington (1968a, 1968b) emphasized the proportionate

but small numbers of minority examinees in norming samples. Their low representation, Harrington (1968a, 1968b) argued, made it impossible for minorities to exert any influence on the results of a test. Harrington devised an innovative experimental test of this proposal.

Harrington (1975, 1976) used six genetically distinct strains of rats to represent ethnicities. He then composed six populations, each with different proportions of the six rat strains. Next, Harrington constructed six intelligence tests resembling Hebb-Williams mazes. These mazes, similar to the Mazes subtest of the Wechsler scales, are commonly used as intelligence tests for rats. Harrington reasoned that tests normed on populations dominated by a given rat strain would yield higher mean scores for that strain.

Groups of rats that were most numerous in a test's norming sample obtained the highest average score on that test. Harrington concluded from additional analyses of the data that a test developed and normed on a White majority could not have equivalent predictive validity for Blacks or any other minority group (1975, 1976).

Reynolds, Lowe et al. (1999) have argued that Harrington's generalizations break down in three respects. Harrington (1975, 1976) interpreted his findings in terms of predictive validity. Most studies have indicated that tests of intelligence and other aptitudes have equivalent predictive validity for racial groups under various circumstances and with many criterion measures.

A second problem noted by Reynolds, Lowe et al. (1999) is that Chinese Americans, Japanese Americans, and Jewish Americans have little representation in the norming samples of most ability tests. According to Harrington's model, they should score low on these tests. However, they score at least as high as Whites on tests of intelligence and of some other aptitudes (Gross, 1967; Marjoribanks, 1972; Tyler, 1965; Willerman, 1979). Jewish and Asian communities notably emphasize education. Thus, it can be hypothesized that there is congruence between the intelligence test and the cultural background of members of these communities. Therefore, their performance on these measures would be higher given that the cultural loading has been minimized (Valencia, Suzuki, & Salinas, 2001).

Finally, Harrington's (1975, 1976) approach can account for group differences in overall test scores but not for patterns of abilities reflected in varying subtest scores. For example, one ethnic group often scores higher than another on some subtests but lower on others. Harrington's model can explain only inequality that is uniform from subtest to subtest. The arguments of Reynolds, Lowe et al. (1999) carry considerable weight, because (a) they are grounded directly in empirical results rather than rational arguments, such as those made by Harrington, and (b) those results have been found with humans; results found with nonhumans cannot be generalized to humans without additional evidence.

Harrington's (1975, 1976) conclusions were overgeneralizations. Rats are simply so different from people that rat and human intelligence cannot be assumed to behave the same. Finally, Harrington used genetic populations in his studies. However, the roles of genetic, environmental, and interactive effects in determining the scores of human ethnic groups are still topics of debate, and an interaction is the preferred explanation. Harrington begged the nature-nurture question, implicitly presupposing heavy genetic effects.

The focus of Harrington's (1975, 1976) work was reduced scores for minority examinees, an important avenue of investigation. Artifactually low scores on an intelligence test could lead to acts of race discrimination, such as misassignment to educational programs or spurious denial of employment. This issue is the one over which most court cases involving test bias have been contested (Reynolds, Lowe, et al., 1999).

## MEAN DIFFERENCES AS TEST BIAS

A view widely held by laypeople and researchers (Adebimpe, Gigandet, & Harris, 1979; Alley & Foster, 1978; Hilliard, 1979, 1984; Jackson, 1975; Mercer, 1976; Padilla, 1988; Williams, 1974; Wright & Isenstein, 1977–1978) is that group differences in mean scores on ability tests constitute test bias. As adherents to this view contend, there is no valid, a priori reason to suppose that cognitive ability should differ from one ethnic group to another. However, the same is true of the assumption that cognitive ability should be the same for all ethnic groups and that any differences shown on a test must therefore be effects of bias. As noted by Reynolds, Lowe et al. (1999), an a priori acceptance of either position is untenable from a scientific standpoint.

Some authors add that the distributions of test scores of each ethnic group, not merely the means, must be identical before one can assume that a test is fair. Identical distributions, like equal means, have limitations involving accuracy. Such alterations correct for any source of score differences, including those for which the test is not

responsible. Equal scores attained in this way necessarily depart from reality to some degree.

### Egalitarian Fallacy

Jensen (1980; Brown et al., 1999) contended that three fallacious assumptions were impeding the scientific study of test bias: (a) the *egalitarian fallacy*, that all groups were equal in the characteristics measured by a test, so that any score difference must result from bias; (b) the *culture-bound fallacy*, that reviewers can assess the culture loadings of items through casual inspection or armchair judgment; and (c) the *standardization fallacy*, that a test is necessarily biased when used with any group not included in large numbers in the norming sample. In Jensen's view, the mean-difference-as-bias approach is an example of the egalitarian fallacy.

A prior assumption of equal ability is as unwarranted scientifically as the opposite assumption. Studies have shown group differences for many abilities and even for sensory capacities (Reynolds, Willson et al., 1999). Both equalities and inequalities must be found *empirically*, that is, through scientific observation. An assumption of equality, if carried out consistently, would have a stultifying effect on research. Torrance (1980) observed that disadvantaged Black children in the United States have sometimes earned higher creativity scores than many White children. This finding may be important, given that Blacks are underrepresented in classes for gifted students. The egalitarian assumption implies that these Black children's high creativity is an artifact of tests, foreclosing on more substantive interpretations—and on possible changes in student placement.

Equal ability on the part of different ethnic groups is not a defensible egalitarian fallacy. A fallacy, as best understood, is an error in judgment or reasoning, but the question of equal ability is an empirical one. By contrast, an *a priori assumption* of either equal or unequal ability can be regarded as fallacious. The assumption of equal ability is most relevant, because it is implicit when any researcher interprets a mean difference as test bias.

The impossibility of proving a null hypothesis is relevant here. Scientists never regard a null hypothesis as proven, because the absence of a counterinstance cannot prove a rule. If 100 studies do not provide a counterinstance, the 101st study may. Likewise, the failure to reject a hypothesis of equality between groups—that is, a null hypothesis—cannot prove that the groups are equal. This hypothesis, then, is not falsifiable and is therefore problematic for researchers.

### Limitations of Mean Differences

As noted, a mean difference by itself does not show bias. One may ask, then, what (if anything) it does show. It indicates simply that two groups differ when means are taken to represent their performance. Thus, its accuracy depends on how well means, as opposed to other measures of the typical score, represent the two groups; on how well *any* measure of the typical score *can* represent the two groups; and on how well *differences* in typical scores, rather than in variation, asymmetry, or other properties, can represent the relationships between the two groups. Ramsay (2000) reanalyzed a study in which mean differences between groups had been found. The reanalysis showed that the two groups differed much more in variation than in typical scores.

Most important, a mean difference provides no information as to *why* two groups differ: because of test bias, genetic influences, environmental factors, a gene-environment interaction, or perhaps biases in society recorded by tests. Rather than answering this question, mean differences raise it in the first place. Thus, they are a starting point—but are they a good one? Answering this question is a logical next step.

A difference between group means is easy to obtain. In addition, it permits an easy, straightforward interpretation—but a deceptive one. It provides scant information, and none at all regarding variation, kurtosis, or asymmetry. These additional properties are needed to understand any group's scores.

Moreover, a mean difference is often an inaccurate measure of center. If a group's scores are highly asymmetric—that is, if the high scores taper off gradually but the low scores clump together, or vice versa—their mean is always too high or too low, pulled as it is toward the scores that taper gradually. Symmetry should never be assumed, even for standardized test scores. A test with a large, national norming sample can produce symmetric scores with that sample but asymmetric or *skewed* scores for particular schools, communities, or geographic regions. Results for people in these areas, if skewed, can produce an inaccurate mean and therefore an inaccurate mean difference. Even a large norming sample can include very small samples for one or more groups, producing misleading mean differences for the norming sample itself.

Finally, a mean is a point estimate: a single number that summarizes the scores of an entire group of people. A group's scores can have little skew or kurtosis but vary so widely that the mean is not typical of the highest and lowest

scores. In addition to being potentially inaccurate, then, a mean can be unrepresentative of the group it purports to summarize.

Thus, means have numerous potential limitations as a way to describe groups and differences between groups. In addition to a mean, measures of shape and spread, sometimes called *distribution* and *variation*, are necessary. Researchers, including clinical researchers, sometimes may need to use different centroids entirely: medians, modes, or modified *M* statistics. Most basically, we always need a thoroughgoing description of each sample. Furthermore, it is both possible and necessary to test the characteristics of each sample to assess their representativeness of the respective population characteristics. This testing can be a simple process, often using group confidence intervals.

Once we know what we have found—which characteristics vary from group to group—we can use this information to start to answer the question *why*. That is, we can begin to investigate causation. Multivariate techniques are often suitable for this work. Bivariate techniques address only two variables, as the name implies. Thus, they are ill suited to pursue possible causal relationships, because they cannot rule out alternative explanations posed by additional variables (Ramsay, 2000).

Alternatively, we can avoid the elusive causal question *why* and instead use measurement techniques developed to assess bias. Reynolds (1982a; Reynolds & Carson, 2005) provides copious information about these techniques. Such procedures cannot tell us if group differences result from genetic or environmental factors, but they can suggest whether test scores may be biased. Researchers have generated a literature of considerable size and sophistication using measurement techniques for examining test bias. This chapter next considers the results of such research.

## RESULTS OF BIAS RESEARCH

Methods of detecting bias include using explicit procedures to determine content validity, oversampling of particular racial and ethnic groups, and employing statistical procedures to address potential concerns. Enhanced computer technology has also enabled implementation of alternative testing formats (e.g., item response theory) and other methods to determine equitable assessment across diverse racial and ethnic groups, taking into consideration testing procedures, scoring, and use of scores (Dana, 2005; Mpofu & Ortiz, 2009).

A review of 62 cultural bias studies conducted by Valencia et al. (2001) determined that most of the studies were conducted in the 1980s, with fewer studies being conducted in the 1990s due to the consistent finding that "prominent intelligence tests" like the WISC/WISC-R were found to be nonbiased. In addition, the studies were "overwhelmingly based on African American and Mexican American children" (p. 120). A substantial proportion of the studies did not control for SES, language dominance and proficiency, and sex of the participants in the bias evaluation. The majority of the studies 71% ($n = 44$) indicated nonbiased results while 29% ($n = 18$) were found to have mixed or biased findings. The next sections provide greater detail regarding the seminal review of test bias studies by Jensen (1980), which provided a major impetus for published research in the years that followed.

### Review by Jensen

Jensen (1980) compiled an extensive early review of test bias studies. One concern addressed in the review was rational judgments that test items were biased based on their content or phrasing. For scientists, *rational* judgments are those based on reason rather than empirical findings. Such judgments may seem sound or even self-evident, but they often conflict with each other and with scientific evidence.

A WISC-R item (also included on the WISC-IV) often challenged on rational grounds is "What is the thing to do if a boy/girl much smaller than yourself starts to fight with you?" Correct responses include "Walk away" and "Don't hit him back." CTBH proponents criticized this item as biased against inner-city Black children, who may be expected to hit back to maintain their status and who may therefore respond incorrectly for cultural reasons. Jensen (1980) reviewed large-sample research indicating that proportionately more Black children than White children responded correctly to this item. Miele (1979), who also researched this item in a large-*N* study, concluded that the item was easier for Blacks than for Whites. As with this item, empirical results often contradict rational judgments.

#### *Predictive and Construct Validity*

Jensen (1980) addressed bias in predictive and construct validity along with situational bias. Bias in predictive validity, as defined by Jensen, is systematic error in predicting a criterion variable for people of different groups. This bias occurs when one regression equation is incorrectly used for two or more groups. The review included studies involving Blacks and Whites, the two most frequently researched groups. The conclusions reached by Jensen were that (a) a large majority of studies showed

that tests were equally valid for these groups and that (b) when differences were found, the tests overpredicted the criterion performance of Black examinees when compared with White examinees. CTBH would have predicted the opposite result.

Bias in construct validity occurs when a test measures groups of examinees differently. For example, a test can be more difficult, valid, or reliable for one group than for another. Construct bias involves the test itself, whereas predictive bias involves a test's prediction of a result outside the test.

Jensen (1980) found numerous studies of bias in construct validity. Regarding difficulty, when item scores differed for ethnic groups or social classes, the differences were not consistently associated with the culture loadings of the tests. Score differences between Black and White examinees were larger on nonverbal than on verbal tests, contrary to beliefs that nonverbal tests are culture fair or unbiased. The sizes of Black–White differences were positively associated with tests' correlations with *g,* or general ability. In tests with several item types, such as traditional intelligence tests, the rank orders of item difficulties for different ethnic groups were very highly correlated. Items that discriminated most between Black and White examinees also discriminated most between older and younger members of each ethnic group. Finally, Blacks, Whites, and Mexican Americans showed similar correlations between raw test scores and chronological ages.

In addition, Jensen (1980) reviewed results pertaining to validity and reliability. Black, White, and Mexican American examinees produced similar estimates of internal consistency reliability. Regarding validity, Black and White samples showed the same factor structures. According to Jensen, the evidence was generally inconclusive for infrequently researched ethnic groups, such as Asian Americans and Native Americans.

### *Situational Bias*

Jensen's (1980) term *situational bias* refers to "influences in the test situation, but independent of the test itself, that may bias test scores" (p. 377). These influences may include, among others, characteristics of the test setting, the instructions, and the examiners themselves. Examples include anxiety, practice and coaching effects, and examiner dialect and ethnic group (Jensen, 1984). As Jensen (1980) observed, situational influences would not constitute test bias, because they are not attributes of the tests themselves. Nevertheless, they should emerge in studies of construct and predictive bias. Jensen concluded that the situational variables reviewed did not influence group differences in scores.

Soon after Jensen's (1980) review was published, the National Academy of Sciences and the National Research Council commissioned a panel of 19 experts, who conducted a second review of the test bias literature. The panel concluded that well-constructed tests were not biased against African Americans or other English-speaking minority groups (Wigdor & Garner, 1982). Later, a panel of 52 professionals signed a position paper that concluded, in part: "Intelligence tests are not culturally biased against American blacks or other native-born, English-speaking peoples in the United States. Rather, IQ scores predict equally accurately for all such Americans, regardless of race and social class" ("Mainstream Science," 1994, p. A18). That same year, a task force of 11 psychologists, established by the APA Association, concluded that no test characteristic reviewed made a substantial contribution to Black–White differences in intelligence scores (Neisser et al., 1996). Thus, several major reviews have failed to support CTBH. (See also Reynolds, 1998a, 1999.)

## Review by Reynolds, Lowe, and Saenz

The next sections highlight the work of Reynolds, Lowe, and Saenz (1999) focusing on content, construct, and predictive validity in relation to issues of bias.

### *Content Validity*

*Content validity* is the extent to which the content of a test is a representative sample of the behavior to be measured (Anastasi, 1988). Items with content bias should behave differently from group to group for people of the same standing on the characteristic being measured. Typically, reviewers judge an intelligence item to have content bias because the information or solution method required is unfamiliar to disadvantaged or minority individuals, or because the test's author has arbitrarily decided on the correct answer, so that minorities are penalized for giving responses that are correct in their own culture but not in the author's culture.

The issue of content validity with achievement tests is complex. Important variables to consider include exposure to instruction, general ability of the group, and accuracy and specificity of the items for the sample (Reynolds, Lowe et al., 1999; see also Schmidt, 1983). Little research is available for personality tests, but cultural variables that may be found to influence some personality tests include beliefs regarding discipline and aggression, values related

to education and employment, and perceptions concerning society's fairness toward one's group.

Camilli and Shepard (1994; Reynolds, 2000a) recommended techniques based on item-response theory to detect differential item functioning (DIF). DIF statistics detect items that behave differently from one group to another. A statistically significant DIF statistic, by itself, does not indicate bias but may lead to later findings of bias through additional research, with consideration of the construct meant to be measured. For example, if an item on a composition test were about medieval history, studies might be conducted to determine if the item is measuring composition skill or some unintended trait, such as historical knowledge. For smaller samples, a contingency table (CT) procedure is often used to estimate DIF. CT approaches are relatively easy to understand and interpret.

Freedle and Kostin (1997) used ethnic comparison to examine factors that may have impacted DIF values on the Scholastic Aptitude Test (SAT) and Graduate Record Exam (GRE) analogy items comparing Black and White examinees matched for total verbal score. African American examinees performed better than Whites on analogy items that had a social-personality content as opposed to a science content. The authors proposed two concepts, cultural familiarity and semantic ambiguity, to explain the persistent pattern of results indicating that Black examinees and other minority groups consistently perform differentially better on harder verbal items and differentially worse on easier items. The "easy" items contain more culturally specific content and can be viewed differently based on cultural and socioeconomic background. The "hard" items do not generally contain words that have variable definitions because they are familiar to those with higher levels of education. Freedle (2003) noted that African American and White examinees disagreed in how they responded to "common" words, such as "valuable," "justice," "progress," and "class" (p. 7). "Such words, when presented in restricted verbal context, can potentially be misinterpreted across racial groups" (Freedle, 2010, p. 396). These findings have been replicated by other scholars, indicating that SAT items function differently for African American and White subgroups (Santelices & Wilson, 2010).

Nandakumar, Glutting, and Oakland (1993) used a CT approach to investigate possible racial, ethnic, and gender bias on the Guide to the Assessment of Test Session Behavior (GATSB). Participants were boys and girls age 6 to 16 years, of White, Black, or Hispanic ethnicity. Only 10 of 80 items produced statistically significant DIFs, suggesting that the GATSB has little bias for different genders and ethnicities.

In very-large-*N* studies, Reynolds, Willson, and Chatman (1984) used a partial correlation procedure (Reynolds, 2000a) to estimate DIF in tests of intelligence and related aptitudes. The researchers found no systematic bias against African Americans or women on measures of English vocabulary. Willson, Nolan, Reynolds, and Kamphaus (1989) used the same procedure to estimate DIF on the Mental Processing scales of the K-ABC. The researchers concluded that there was little apparent evidence of race or gender bias.

Jensen (1976) used a chi-square technique (Reynolds, 2000a) to examine the distribution of incorrect responses for two multiple-choice intelligence tests, RPM and the Peabody Picture-Vocabulary Test (PPVT). Participants were Black and White children age 6 to 12 years. The errors for many items were distributed systematically over the response options. This pattern, however, was the same for Blacks and Whites. These results indicated bias in a general sense, but not racial bias. On RPM, Black and White children made different types of errors, but for few items. The researcher examined these items with children of different ages. For each of the items, Jensen was able to duplicate Blacks' response patterns using those of Whites approximately 2 years younger.

Scheuneman (1987) used linear methodology on GRE item data to show possible influences on the scores of Black and White test takers. Vocabulary content, true-false response, and presence or absence of diagrams were among the item characteristics examined. Paired, experimental items were administered in the experimental section of the GRE General Test, given in December 1982. Results indicated that certain characteristics common to a variety of items may have a differential influence on Blacks' and Whites' scores. These items may be measuring, in part, test content rather than verbal, quantitative, or analytical skill.

Jensen (1974, 1976, 1977) evaluated bias on the Wonderlic Personnel Test (WPT), PPVT, and RPM using correlations between *P* decrements (Reynolds, 2000a) obtained by Black students and those obtained by White students. *P* is the probability of passing an item, and a *P* decrement is the size of the difference between *P*s for one item and the next. Thus, *P*-decrements represent an "analysis of the ordering of the difficulties of the items as a whole, and the degree to which such ordering remains constant across groups" (Reynolds & Carson, 2005, p. 803). Jensen also obtained correlations between the rank orders of item difficulties for Black and Whites. Results for rank orders and

*P* decrements, it should be noted, differ from those that would be obtained for the scores themselves.

The tests examined were RPM; the PPVT; the WISC-R; the WPT; and the Revised Stanford-Binet Intelligence Scale, Form L-M. Jensen (1974) obtained the same data for Mexican American and White students on the PPVT and RPM. Table 4.1 shows the results, with similar findings obtained by Sandoval (1979) and Miele (1979). The correlations showed little evidence of content bias in the scales examined. Most correlations appeared large. Some individual items were identified as biased, but they accounted for only 2% to 5% of the variation in score differences.

Hammill (1991) used correlations of *p* decrements to examine the Detroit Tests of Learning Aptitude (DTLA-3). Correlations exceeded .90 for all subtests, and most exceeded .95. Reynolds and Bigler (1994) presented correlations of *P* decrements for the 14 subtests of the Test of Memory and Learning (TOMAL). Correlations again exceeded .90, with most exceeding .95, for males and females and for all ethnicities studied.

Another procedure for detecting item bias relies on the partial correlation between an item score and a nominal variable, such as ethnic group. The correlation partialed out is that between total test score and the nominal variable. If the variable and the item score are correlated after the partialed correlation is removed, the item is performing differently from group to group, which suggests bias. Reynolds, Lowe et al. (1999) described this technique as a powerful means of detecting item bias. They noted, however, that it is a relatively recent application. Thus, it may have limitations not yet known.

Research on item bias in personality measures is sparse but has produced results similar to those with ability tests (Moran, 1990; Reynolds, 1998a, 1998b; Reynolds & Harding, 1983). The few studies of behavior rating scales have produced little evidence of bias for White, Black, and Hispanic and Latin populations in the United States (James, 1995; Mayfield & Reynolds, 1998; Reynolds & Kamphaus, 1992).

Not all studies of content bias have focused on items. Researchers evaluating the WISC-R have defined bias differently. Few results are available for the WISC-III; future research should use data from this newer test. Prifitera and Saklofske (1998) addressed the WISC-III and ethnic bias in the United States. These results are discussed later in the "Construct Validity" and "Predictive Validity" sections.

Reynolds and Jensen (1983) examined the 12 WISC-R subtests for bias against Black children using a variation of the group by item analysis of variance (ANOVA). The researchers matched Black children to White children from the norming sample on the basis of gender and Full Scale IQ. SES was a third matching variable and was used when a child had more than one match in the other group. Matching controlled for *g*, so a group difference indicated that the subtest in question was more difficult for Blacks or for Whites.

Black children exceeded White children on Digit Span and Coding. Whites exceeded Blacks on Comprehension, Object Assembly, and Mazes. Blacks tended to obtain higher scores on Arithmetic and Whites on Picture Arrangement. The actual differences were very small, and variance due to ethnic group was less than 5% for each subtest. If the WISC-R is viewed as a test measuring only *g,* these results may be interpretable as indicating subtest bias. Alternatively, the results may indicate differences in Level II ability (Reynolds, Willson et al., 1999) or in specific or intermediate abilities.

Taken together, studies of major ability and personality tests show no consistent evidence for content bias. When

**TABLE 4.1 Ethnic Correlations for *P* Decrements and for Rank Orders of Item Difficulties**

| | Black–White | | | | Mexican American–White | | | |
|---|---|---|---|---|---|---|---|---|
| Scale | Rank Orders | | *P* Decrements | | Rank Orders | | *P* Decrements | |
| PPVT (Jensen, 1974) | .99[a] | .98[b] | .79[a] | .65[b] | .98[a] | .98[b] | .78[a] | .66[b] |
| RPM (Jensen, 1974) | .99[a] | .99[b] | .98[a] | .96[b] | .99[a] | .99[b] | .99[a] | .97[b] |
| SB L-M (Jensen, 1976) | .96[c] | | | | | | | |
| WISC-R (Jensen, 1976) | .95[c] | | | | | | | |
| (Sandoval, 1979) | .98[c] | | .87[c] | | | .99[c] | | .91[c] |
| WISC (Miele, 1979) | .96[a] | .95[b] | | | | | | |
| WPT (Jensen, 1977) | .94[c] | | .81[c] | | | | | |

*Notes.* PPVT = Peabody Picture Vocabulary Test; RPM = Raven's Progressive Matrices; SB L-M = Stanford-Binet, Form LM; WISC-R = Wechsler Intelligence Scale for Children-Revised; WPT= Wonderlic Personnel Test; Sandoval, 1979 = Medians for 10 WISC-R subtests, excluding Coding and Digit Span.

[a] Males.

[b] Females.

[c] Males and females combined.

bias is found, it is small. Tests with satisfactory reliability, validity, and norming appear also to have little content bias. For numerous standardized tests, however, results are not yet available. Research with these tests should continue investigating possible content bias with differing ethnic and other groups.

## Construct Validity

Anastasi (1988) has defined construct validity as the extent to which a test may be said to measure a theoretical construct or trait. Test bias in construct validity, then, may be defined as the extent to which a test measures different constructs for different groups.

Factor analysis is a widely used method for investigating construct bias (Reynolds, 2000a). This set of complex techniques groups together items or subtests that correlate highly among themselves. When a group of items correlates highly together, the researcher interprets them as reflecting a single characteristic. The researcher then examines the pattern of correlations and induces the nature of this characteristic. Table 4.2 shows a simple example.

In the table, the subtests picture identification, matrix comparison, visual search, and diagram drawing have high correlations in the column labeled "Factor 1." Definitions, antonyms, synonyms, and multiple meanings have low correlations in this column but much higher ones in the column labeled "Factor 2." A researcher might interpret these results as indicating that the first four subtests correlate with factor 1 and the second four correlate with factor 2. Examining the table, the researcher might see that the subtests correlating highly with factor 1 require visual activity, and he or she might therefore label this factor Visual Ability. The same researcher might see that the subtests correlating highly with factor 2 involve the meanings of words, and he or she might label this factor Word Meanings. To label factors in this way, researchers must be familiar with the subtests or items, common responses to them, and scoring of these responses. (See also Ramsay & Reynolds, 2000a.) The results in Table 4.2 are called a *two-factor solution*. Actual factor analysis is a set of advanced statistical techniques, and the explanation presented here is necessarily a gross oversimplification.

**TABLE 4.2 Sample Factor Structure**

| Subtest | Factor 1 | Factor 2 |
|---|---|---|
| Picture Identification | *.78* | .17 |
| Matrix Comparison | *.82* | .26 |
| Visual Search | *.86* | .30 |
| Diagram Drawing | *.91* | .29 |
| Definitions | .23 | *.87* |
| Antonyms | .07 | *.92* |
| Synonyms | .21 | *.88* |
| Multiple Meanings | .36 | *.94* |

Very similar factor analytic results for two or more groups, such as genders or ethnicities, are evidence that the test responses being analyzed behave similarly as to the constructs they represent and the extent to which they represent them. As noted by Reynolds, Lowe et al. (1999), such comparative factor analyses with multiple populations are important for the work of clinicians, who must know that a test functions very similarly from one population to another to interpret scores consistently.

Researchers most often calculate a coefficient of congruence or simply a Pearson correlation to examine factorial similarity, often called *factor congruence* or *factor invariance*. The variables correlated are one group's item or subtest correlations (shown in Table 4.2) with another's. A coefficient of congruence may be preferable, but the commonly used techniques produce very similar results, at least with large samples (Reynolds & Harding, 1983; Reynolds, Lowe et al., 1999). Researchers frequently interpret a value of .90 or higher as indicating factor congruity. For other applicable techniques, see Reynolds (2000a).

Extensive research regarding racial and ethnic groups is available for the widely used WISC and WISC-R. This work consists largely of factor analyses. Psychometricians are trained in this method, so its usefulness in assessing bias is opportune. Unfortunately, many reports of this research fail to specify whether exploratory or confirmatory factor analysis has been used. In factor analyses of construct and other bias, exploratory techniques are most common. Results with the WISC and WISC-R generally support factor congruity. For preschool-age children also, factor analytic results support congruity for racial and ethnic groups (Reynolds, 1982a).

Reschly (1978) conducted factor analyses comparing WISC-R correlations for Blacks, Whites, Mexican Americans, and Papagos, a Native American group, all in the southwestern United States. Reschly found that the two-factor solutions were congruent for the four ethnicities. The 12 coefficients of congruence ranged from .97 to .99. For the less widely used three-factor solutions, only results for Whites and Mexican Americans were congruent. The one-factor solution showed congruence for all four ethnicities, as Miele (1979) had found with the WISC.

Oakland and Feigenbaum (1979) factor-analyzed the 12 WISC-R subtests separately for random samples of normal Black, White, and Mexican American children from an urban school district in the northwestern United States.

Samples were stratified by race, age, sex, and SES. The researchers used a Pearson *r* for each factor to compare it for the three ethnic groups. The one-factor solution produced *r*s of .95 for Black and White children, .97 for Mexican American and White children, and .96 for Black and Mexican American children. The remaining results were $r = .94$ to .99. Thus, WISC-R scores were congruent for the three ethnic groups.

Gutkin and Reynolds (1981) compared factor analytic results for the Black and White children in the WISC-R norming sample. Samples were stratified by age, sex, race, SES, geographic region, and community size to match 1970 U.S. Census Bureau data. The researchers compared one-, two-, and three-factor solutions using magnitudes of unique variances, proportion of total variance accounted for by common factor variance, patterns of correlations with each factor, and percentage of common factor variance accounted for by each factor. Coefficients of congruence were .99 for comparisons of the unique variances and of the three solutions examined. Thus, the factor correlations were congruent for Black and White children.

Dean (1979) compared three-factor WISC-R solutions for White and Mexican American children referred because of learning difficulties in the regular classroom. Analyzing the 10 main WISC-R subtests, Dean found these coefficients of congruence: .84 for Verbal Comprehension, .89 for Perceptual Organization, and .88 for Freedom from Distractibility.

Gutkin and Reynolds (1980) compared one-, two-, and three-factor principal-factor solutions of the WISC-R for referred White and Mexican American children. The researchers also compared their solutions to those of Reschly (1978) and to those derived from the norming sample. Coefficients of congruence were .99 for Gutkin and Reynolds's one-factor solutions and .98 and .91 for their two-factor solutions. Coefficients of congruence exceeded .90 in all comparisons of Gutkin and Reynolds's solutions to Reschly's solutions for normal Black, White, Mexican American, and Papago children and to solutions derived from the norming sample. Three-factor results were more varied but also indicated substantial congruity for these children.

DeFries et al. (1974) administered 15 ability tests to large samples of American children of Chinese or Japanese ancestry. The researchers examined correlations among the 15 tests for the two ethnic groups and concluded that the cognitive organization of the groups was virtually identical. Willerman (1979) reviewed these results and concluded, in part, that the tests were measuring the same abilities for the two groups of children.

Results with adults are available as well. Kaiser (1986) and Scholwinski (1985) have found the Wechsler Adult Intelligence Scale–Revised (WAIS-R) to be factorially congruent for Black and White adults from the norming sample. Kaiser conducted separate hierarchical analyses for Black and White participants and calculated coefficients of congruence for the General, Verbal, and Performance factors. Coefficients for the three factors were .99, .98, and .97, respectively. Scholwinski selected Black and White participants closely matched in age, sex, and Full Scale IQ, from the WAIS-R norming sample. Results again indicated factorial congruence.

Edwards and Oakland (2006) examined the factorial invariance of Woodcock-Johnson III (WJ-III) scores for African Americans and Caucasian American students in the norming sample. Results indicate that although their mean scores differ, the WJ-III scores have comparable meaning across groups, as evidenced by the consistent factor structure found for both groups.

Researchers have also assessed construct bias by estimating internal consistency reliabilities for different groups. *Internal consistency reliability* is the extent to which all items of a test are measuring the same construct. A test is unbiased with regard to this characteristic to the extent that its reliabilities are similar from group to group.

Jensen (1977) used Kuder-Richardson formula 21 to estimate internal consistency reliability for Black and White adults on the Wonderlic Personnel Test. Reliability estimates were .86 and .88 for Blacks and Whites, respectively. In addition, Jensen (1974) used Hoyt's formula to obtain internal consistency estimates of .96 on the PPVT for Black, White, and Mexican American children. The researcher then subdivided each group of children by gender and obtained reliabilities of .95 to .97. Raven's colored matrices produced internal consistency reliabilities of .86 to .91 for the same six race–gender groupings. For these three widely used aptitude tests, Jensen's (1974, 1976) results indicated homogeneity of test content and consistency of measurement by gender and ethnicity.

Sandoval (1979) and Oakland and Feigenbaum (1979) have extensively examined the internal consistency reliability of the WISC-R subtests, excluding Digit Span and Coding, for which internal consistency analysis is inappropriate. Both studies included Black, White, and Mexican American children. Both samples were large, with Sandoval's exceeding 1,000.

Sandoval (1979) estimated reliabilities to be within .04 of each other for all subtests except Object Assembly. This subtest was most reliable for Black children at .95, followed by Whites at .79 and Mexican Americans at .75.

Oakland and Feigenbaum (1979) found reliabilities within .06, again excepting Object Assembly. In this study, the subtest was most reliable for Whites at .76, followed by Blacks at .64 and Mexican Americans at .67. Oakland and Feigenbaum also found consistent reliabilities for males and females.

Dean (1979) assessed the internal consistency reliability of the WISC-R for Mexican American children tested by White examiners. Reliabilities were consistent with, although slightly larger than, those reported by Wechsler (1975) for the norming sample.

Results with the WISC-III norming sample (Prifitera, Weiss, & Saklofske, 1998) suggested a substantial association between IQ and SES. WISC-III Full Scale IQ was higher for children whose parents had high education levels, and parental education is considered a good measure of SES. The children's Full Scale IQs were 110.7, 103.0, 97.9, 90.6, and 87.7, respectively, in the direction of highest (college or above) to lowest (<8th grade) parental education level. Researchers have reported similar results for other IQ tests (Prifitera et al., 1998). Such results should not be taken as showing SES bias because, like ethnic and gender differences, they may reflect real distinctions, perhaps influenced by social and economic factors. Indeed, IQ is thought to be associated with SES. By reflecting this theoretical characteristic of intelligence, SES differences may support the construct validity of the tests examined.

Psychologists view intelligence as a developmental phenomenon (Reynolds, Lowe et al., 1999). Hence, similar correlations of raw scores with age may be evidence of construct validity for intelligence tests. Jensen (1976) found that these correlations for the PPVT were .73 with Blacks, .79 with Whites, and .67 with Mexican Americans. For Raven's colored matrices, correlations were .66 for Blacks, .72 for Whites, and .70 for Mexican Americans. The K-ABC produced similar results (Kamphaus & Reynolds, 1987).

A review by Moran (1990) and a literature search by Reynolds, Lowe et al. (1999) indicated that few construct bias studies of personality tests had been published. This limitation is notable, given large mean differences on the Minnesota Multiphasic Personality Inventory (MMPI), and possibly the MMPI-2. The MMPI is the most widely used and researched personality test in the world (e.g., Butcher, 2009). Patterns of score differences have been noted based on gender and ethnicity (Reynolds, Lowe et al.). In addition, to challenges of cultural bias, Groth-Marnat (2009) noted that score differences may reflect different personality traits, cultural beliefs, and experiences of racial discrimination (e.g., anger, frustration). Other popular measures, such as the Revised Children's Manifest Anxiety Scale (RCMAS), suggest consistent results by gender and ethnicity (Moran, 1990; Reynolds & Paget, 1981).

To summarize, studies using different samples, methodologies, and definitions of bias indicate that many prominent standardized tests are consistent from one race, ethnicity, and gender to another. (See Reynolds, 1982b, for a review of methodologies.) These tests appear to be reasonably unbiased for the groups investigated.

## Predictive Validity

As the term implies, *predictive validity* pertains to *prediction* from test scores, whereas *content* and *construct validity* pertain to *measurement*. Anastasi (1988) defined predictive or criterion-related validity as "the effectiveness of a test in predicting an individual's performance in specified activities" (p. 145). Thus, test bias in predictive validity may be defined as systematic error that affects examinees' performance differentially depending on their group membership. Cleary et al. (1975) defined predictive test bias as constant error in an inference or prediction, or error in a prediction that exceeds the smallest feasible random error, as a function of membership in a particular group. Oakland and Matuszek (1977) found that fewer children were wrongly placed using these criteria than using other, varied models of bias. An early court ruling also favored Cleary's definition (*Cortez v. Rosen*, 1975).

Of importance, inaccurate prediction sometimes reflects inconsistent measurement of the characteristic being predicted rather than bias in the test used to predict it. In addition, numerous investigations of predictive bias have addressed the selection of employment and college applicants of different racial and ethnic groups. Studies also address prediction bias in personality tests (Moran, 1990; Monnot, Quirk, Hoerger, & Brewer, 2009). As the chapter shows, copious results for intelligence tests are available.

Under the definition presented by Cleary et al. (1975), the regression line formed by any predictor and criterion (e.g., total test score and a predicted characteristic) must be the same for each group with whom the test is used. A regression line consists of two parameters: a slope, $a$, and an intercept, $b$. Too great a group difference in either of these parameters indicates that a regression equation based on the combined groups would predict inaccurately (Reynolds, Lowe et al., 1999). A separate equation for each group then becomes necessary with the groups and characteristics for which bias has been found.

Hunter, Schmidt, and Hunter (1979) reviewed 39 studies, yielding 866 comparisons, of Black–White test score validity in personnel selection. The researchers concluded that the results did not support a hypothesis of differential or single-group validity. Several studies of the SAT indicated no predictive bias, or small bias against Whites, in predicting grade point average and other measures of college performance (Cleary, 1968; Cleary et al., 1975).

Reschly and Sabers (1979) examined the validity of WISC-R IQs in predicting the Reading and Math subtest scores of Blacks, Whites, Mexican Americans, and Papago Native Americans on the Metropolitan Achievement Tests (MAT). The MAT has undergone item analysis procedures to eliminate content bias, making it especially appropriate for this research: Content bias can be largely ruled out as a competing explanation for any invalidity in prediction. WISC-R IQs underpredicted MAT scores for Whites compared with the remaining groups. Overprediction was greatest for Papagos. The intercept typically showed little bias.

Reynolds and Gutkin (1980) conducted similar analyses for WISC-R Verbal, Performance, and Full Scale IQs as predictors of arithmetic, reading, and spelling. The samples were large groups of White and Mexican American children from the southwestern United States. Only the equation for Performance IQ and arithmetic achievement differed for the two groups. Here an intercept bias favored Mexican American children.

Likewise, Reynolds and Hartlage (1979) assessed WISC and WISC-R Full Scale IQs as predictors of Blacks' and Whites' arithmetic and reading achievement. The children's teachers had referred them for psychological services in a rural, southern school district. The researchers found no statistically significant differences for these children. Many participants, however, had incomplete data (34% of the total).

Prifitera et al. (1998) noted studies in which the WISC-III predicted achievement equally for Black, White, and Hispanic children. In one study, Weiss and Prifitera (1995) examined WISC-III Full Scale IQ as a predictor of Wechsler Individual Achievement Test (WIAT) scores for Black, White, and Hispanic children age 6 to 16 years. Results indicated little evidence of slope or intercept bias, a finding consistent with those for the WISC and WISC-R. Weiss, Prifitera, and Roid (1993) reported similar results.

Bossard, Reynolds, and Gutkin (1980) analyzed the 1972 Stanford-Binet Intelligence Scale when used to predict the reading, spelling, and arithmetic attainment of referred Black and White children. No statistically significant bias appeared in comparisons of either correlations or regression analyses.

Reynolds, Willson, and Chatman (1985) evaluated K-ABC scores as predictors of Black and White children's academic attainment. Some of the results indicated bias, usually overprediction of Black children's attainment. Of 56 Potthoff comparisons (i.e., determining bias based on whether the regression equation relating two variables is constant across groups; Reynolds, 1982), however, most indicated no statistically significant bias. Thus, evidence for bias had low method reliability for these children.

In addition, Kamphaus and Reynolds (1987) reviewed seven studies on predictive bias with the K-ABC. Overprediction of Black children's scores was more common than with other tests and was particularly common with the Sequential Processing Scale. The differences were small and were mitigated by using the K-ABC Mental Processing Composite. Some underprediction of Black children's scores also occurred.

A series of very-large-*N* studies reviewed by Jensen (1980) and Sattler (1974) compared the predictive validities of group IQ tests for different races. This procedure has an important limitation. If validities differ, regression analyses must also differ. If validities are the same, regression analyses may nonetheless differ, making additional analysis necessary (but see Reynolds, Lowe, et al., 1999). In addition, Jensen and Sattler found few available studies that followed this method of analysis on which to base their results. Lorge-Thorndike Verbal and Nonverbal IQs were the results most often investigated. The reviewers concluded that validities were comparable for Black and White elementary school children. Despite the fact that decades have passed since the publications by Jensen and Sattler, there still exists the need for researchers to broaden the range of group intelligence tests that they examine. Emphasis on a small subset of available measures continues to be a common limitation of test research.

Guterman (1979) reported an extensive analysis of the Ammons and Ammons Quick Test (QT), a verbal IQ measure, with adolescents of different social classes. The variables predicted were (a) social knowledge measures; (b) school grades obtained in Grades 9, 10, and 12; (c) Reading Comprehension Test scores on the Gates Reading Survey; and (d) Vocabulary and Arithmetic subtest scores on the General Aptitude Test Battery (GATB). Guterman found little evidence of slope or intercept bias with these adolescents, except one social knowledge measure, sexual knowledge, showed intercept bias.

Another extensive analysis merits attention, given its unexpected results. Reynolds (1978) examined seven major

preschool tests: the Draw-a-Design and Draw-a-Child subtests of the McCarthy Scales, the Mathematics and Language subtests of the Tests of Basic Experiences, the Preschool Inventory–Revised Edition, and the Lee-Clark Readiness Test. Variables predicted were four MAT subtests: Word Knowledge, Word Discrimination, Reading, and Arithmetic. Besides increased content validity, the MAT had the advantage of being chosen by teachers in the district as the test most nearly measuring what was taught in their classrooms. Reynolds compared correlations and regression analyses for the following race-gender combinations: Black females versus Black males, White females versus White males, Black females versus White females, and Black males versus White males. The result was 112 comparisons each for correlations and regression analyses.

For each criterion, scores fell in the same rank order: White females < White males < Black females < Black males. Mean validities comparing pre- and posttest scores, with 12 months intervening, were .59 for White females, .50 for White males, .43 for Black females, and .30 for Black males. In spite of these overall differences, only three differences between correlations were statistically significant, a chance finding with 112 comparisons. Potthoff comparisons of regression lines, however, indicated 43 statistically significant differences. Most of these results occurred when race rather than gender was compared: 31 of 46 comparisons ($p < .01$). The Preschool Inventory and Lee-Clark Test most frequently showed bias; the Metropolitan Readiness Tests (MRT) never did. The observed bias overpredicted scores of Black and male children.

Researchers should investigate possible reasons for these results, which may have differed for the seven predictors but also by the statistical results compared. Either Potthoff comparisons or comparisons of correlations may be inaccurate or inconsistent as analyses of predictive test bias. (See also Reynolds, 1980.)

Brief screening measures tend to have low reliability compared with major ability and aptitude tests such as the WISC-III and the K-ABC. Low reliability can lead to bias in prediction (Reynolds, Lowe et al., 1999). More reliable measures, such as the MRT, the WPPSI, and the McCarthy Scales, have shown little evidence of internal bias. The WPPSI and McCarthy Scales have not been assessed for predictive bias with differing racial or ethnic groups (Reynolds, Lowe et al., 1999).

Reynolds (1980) examined test and subtest scores for the seven tests noted earlier when used to predict MAT scores for males and females and for diverse ethnic groups. The researcher examined *residuals*—the differences between predicted scores and actual scores obtained by examinees. Techniques used were multiple regression to obtain residuals and race by gender ANOVA to analyze them.

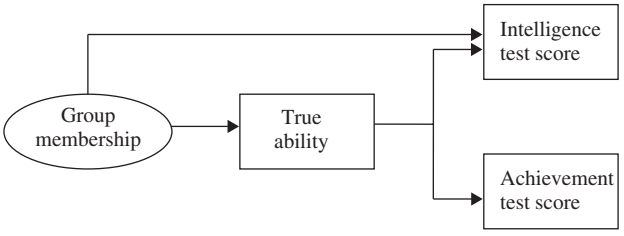
ANOVA results indicated no statistically significant differences in residuals for ethnicities or genders and no statistically significant interactions. Reynolds (1980) then examined a subset of the seven-test battery. No evidence of racial bias appeared. The results indicated gender bias in predicting two of the four MAT subtests, Word Discrimination and Word Knowledge. The seven tests consistently underpredicted females' scores. The difference was small, on the order of .13 to .16 *SD*.

Bias studies are especially critical on personality tests used to diagnose mental disorders, although much of the research conducted on the most popular personality tests (e.g., MMPI-2) continues to focus on differences in scoring patterns between racial and ethnic groups (Suzuki, Onoue, & Hill, forthcoming). A study by Monnot et al. (2009) on male veteran inpatients revealed a "modest" pattern of predictive bias across numerous scales. The authors concluded that "[t]hese biases indicate both over- and underprediction of psychiatric disorders among African Americans on a variety of scales suggesting differential accuracy for the MMPI-2 in predicting diagnostic status between subgroups of male veteran inpatients seeking substance abuse treatment" (p. 145). By comparison, no evidence of overprediction of diagnosis was found for Caucasians across the test scores.

For predictive validity, as for content and construct validity, the results reviewed suggest little evidence of bias, whether differential or single-group validity. Differences are infrequent. Where they exist, they usually take the form of small overpredictions for lower-scoring groups, such as disadvantaged, low-SES, or ethnic minority examinees. These overpredictions are unlikely to account for adverse placement or diagnosis of these groups. On a grander scale, the small differences found may be reflections, but would not be major causes, of sweeping social inequalities affecting ethnic group members. The causes of such problems as employment discrimination and economic deprivation lie primarily outside the testing environment.

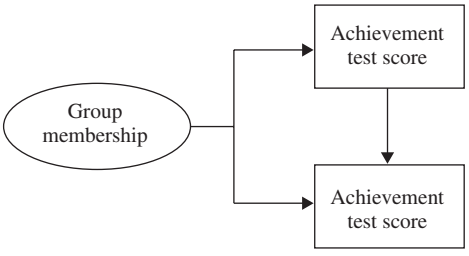
### Path Modeling and Predictive Bias

Keith and Reynolds (1990; see also Ramsay, 1997) have suggested path analysis as a means of assessing predictive bias. Figure 4.1 shows one of their models. Each arrow represents a path, and each oblong or rectangle represents a variable.




The arrow from group membership to intelligence test score represents bias

**Figure 4.1** Path model showing predictive bias




The arrow from group membership to predictor of school achievement represents bias

**Figure 4.2** Revised path model showing predictive bias

The path from group membership to intelligence test score denotes bias. Its beta value, then, should be small. The absence of this path would represent bias of zero.

A limitation of this approach is that no true ability measures exist. Thus, a path model could not incorporate true ability unless it was measured by three or more existing variables. Figure 4.2 shows a proposed model that disposes of this limitation. Here true ability drops out, and a path leads from the predictor, *Achievement Test Score,* to the criterion, *School Achievement*. The path from group membership to the predictor denotes bias; as before, its beta value should be small. The absence of this path would, again, reflect zero bias.

## EXAMINER–EXAMINEE RELATIONSHIP

Contrary findings notwithstanding, many psychological professionals continue to assert that White examiners impede the test performance of minority group members (Sattler, 1988). Sattler and Gwynne (1982) reviewed 27 published studies on the effects of examiners' race on the test scores of children and youth on a wide range of cognitive tests. Participants were students in preschool through Grade 12, most from urban areas throughout the United States. Tests included the Wechsler Scales; the Stanford-Binet, Form L-M; the PPVT; the Draw-a-Man Test; the Iowa Test of Preschool Development; and others. In 23 of these studies, examiner's race (Black or White) and test scores of racial groups (Black or White) had no statistically significant association. Sattler and Gwynne reported that the remaining four studies had methodological limitations, including inappropriate statistical tests and designs. Design limitations included lack of a comparison group and of external criteria to evaluate the validity of procedures used.

The question of possible examiner–examinee effects has taken numerous forms. Minority examinees might obtain reduced scores because of their *responses* to examiner–examinee differences. An examiner of a different race, for example, might evoke anxiety or fear in minority children. Research has lent little support to this possibility. Kaufman (1994), for example, found that Black populations obtained their highest scores on tests most sensitive to anxiety.

White examiners may be less effective than Hispanic American examiners when testing Hispanic American children and adolescents. This proposal, too, has received little support. Gerkin (1978) found that examiner's ethnicity (White or Hispanic American) and examiner's bilingual ability (monolingual or bilingual) had no statistically significant association with the WPPSI IQs or the Leiter International Performance Scale scores of children age 4, 5, and 6 years. Morales and George (1976) found that Hispanic bilingual children in Grades 1 to 3 obtained higher WISC-R scores with monolingual non-Hispanic examiners than with bilingual Hispanic examiners, who tested the children in both Spanish and English (Sattler, 1988; Reynolds, Lowe et al., 1999).

These findings suggest that examiner ethnicity has little adverse effect on minority scores. Examiners need to be well trained and competent, however, in administering standardized tests to diverse minority group members. Rapport may be especially crucial for minority examinees, and approaches that are effective with one ethnic group may be less so with another. The *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists* adopted by the American Psychological Association (2002) noted that cultural competence in assessment requires multicultural understanding in the establishment of the relationship with the client and attending to potential measurement limitations including issues of test bias, fairness, and cultural equivalence. As usual, research in this area should continue. Neither researchers nor clinicians can assume that the results reviewed in this chapter typify all future results.

## HELMS AND CULTURAL EQUIVALENCE

As noted, Helms (1992) and other authors have reframed the CTBH approach over time. Helms has addressed the implicit biological and environmental philosophical perspectives used to explain racial and ethnic group differences in tested cognitive ability. Helms's position is that these perspectives stem from inadequate notions of culture and that neither perspective provides useful information about the cultural equivalence of tests for diverse ethnic groups. Assessment of cultural equivalence is necessary to account for minority groups' cultural, social, and cognitive differences from the majority. Helms (2006) noted the complexity of understanding the internalized racial and cultural experiences and environmental socialization that can impact test performance that are unrelated to intelligence and therefore comprise error. These factors may have a greater impact on the test performance of members of racial and ethnic minority groups in comparison to nonminority group members.

For Helms (1992), cultural equivalence should take seven forms (Butcher, 1982): (1) *functional equivalence*, the extent to which test scores have the same meaning for different cultural groups; (2) *conceptual equivalence*, whether test items have the same meaning and familiarity in different groups; (3) *linguistic equivalence*, whether tests have the same linguistic meaning to different groups; (4) *psychometric equivalence*, the extent to which tests measure the same thing for different groups; (5) *testing condition e*quivalence, whether groups are equally familiar with testing procedures and view testing as a means of assessing ability; (6) *contextual equivalence*, the extent to which a cognitive ability is assessed similarly in different contexts in which people behave; and (7) *sampling equivalence*, whether comparable samples of each cultural group are available at the test development, validation, and interpretation stages.

Helms (1992) argued for the diversification of existing tests, the development of new standardized tests, and the formation of explicit principles, hypotheses, assumptions, and theoretical models for investigating cultural differences. In addition, Helms argued that existing frameworks—biological, environmental, and cultural—should be operationally defined.

For future research, Helms (1992) recommended (a) development of measures for determining interracial cultural dependence and levels of acculturation and assimilation in test items, (b) modification of test content to include items that reflect cultural diversity, (c) examination of incorrect responses, (d) incorporation of cognitive psychology into interactive modes of assessment, (e) use of theories to examine environmental content of criteria, and (f) separate racial group norms for existing tests. Researchers should interpret test scores cautiously, Helms suggested, until psychometricians develop more diverse procedures to address cultural concerns.

Helms's (1992) approach, or one like it, is likely to become a future trend. As observed by Reynolds, Lowe et al. (1999), however, much of the work recommended by Helms has been well under way for several decades. (For an extensive treatment, see Cronbach & Drenth, 1972; see also Hambleton, 1994; Van de Vijver & Hambleton, 1996.) Reynolds et al. contended that Helms coined new terms for old constructs and dismissed many studies already addressing the issues she raises. At best, they believe, Helms organized and continues to call attention to long-recognized empirical issues. She emphasized that test takers of color (i.e., African American, Latino/Latina, Asian American, and Native American) "are competing with White test takers whose racial socialization experiences are either irrelevant to their test performance or give them an undeserved advantage" (Helms, 2006, p. 855).

## TRANSLATION AND CULTURAL TESTING

The findings already reviewed do not apply to translations of tests. Use of a test in a new linguistic culture requires that it be redeveloped from the start. One reason for the early success of the Stanford-Binet Intelligence Scale was that Terman reconceptualized it for the United States, reexamining Binet's theory of intelligence, writing and testing new items, and renorming the scales (Reynolds, Lowe et al., 1999).

Terman's work was an exception to a rule of simple translation of the Binet Scales. Even today, few researchers are experienced in procedures for adapting tests and establishing score equivalence. Nonetheless, the procedures are available, and they increase the validity of the adapted tests (Hambleton & Kanjee, 1995). Adaptation of educational and psychological tests most frequently occurs for one of three reasons: to facilitate comparative ethnic studies, to allow individuals to be tested in their own language, or to reduce the time and cost of developing new tests.

Test adaptation has been commonplace for more than 90 years, but the field of cross-cultural and cross-national comparisons is relatively recent. This field has focused on development and use of adaptation guidelines (Hambleton, 1994), ways to interpret and use cross-cultural and cross-national data (Hambleton & Kanjee, 1995;

Poortinga & Malpass, 1986), and especially procedures for establishing item equivalence (Ellis, 1991; Hambleton, 1993; Poortinga, 1983; van de Vijver & Poortinga, 1991). Test items are said to be equivalent when members of each linguistic or cultural group who have the same standing on the construct measured by the tests have the same probability of selecting the correct item response.

A number of test adaptations can be based on the area of equivalence being addressed (i.e., conceptual, cultural, linguistic, or measurement) (van de Vijver & Leung, 2011). For example, cultural adaptations can be made in terms of terminological/factual driven, including accommodation of specific cultural or country characteristics, or norm driven, taking into consideration norms, values and practices. Language adaptations can be linguistic driven to accommodate structural differences in the language or pragmatics driven to address conventional language usage. Van de Vijver and Tanzer (2004) suggested the committee approach in which "[a] group of people, often with different areas of expertise (such as cultural, linguistic, and psychological) prepare a translation" (p. 123). The cooperative effort among this group improves the quality of the translation. Similarly, Geisinger (2005) outlined translation methods that incorporate a team of culturally sensitive translators "who not only translate the assessment device linguistically, but from a cultural perspective as well." This work is then evaluated by a "panel of others who are knowledgeable about the content covered by the assessment, fluent in both the original and target languages, and thoroughly experienced in the two cultures" (p. 197).

The designs used to establish item equivalence fall into two categories, judgmental and statistical. Judgmental designs rely on a person's or group's decision regarding the degree of translation equivalence of an item. Two common designs are forward translation and back translation (Hambleton & Bollwark, 1991). In the first design, translators adapt or translate a test to the target culture or language. Other translators then assess the equivalency of the two versions. If the versions are not equivalent, changes are made. In the second design, translators adapt or translate a test to the target culture or language as before. Other translators readapt the items back to the original culture or language. An assessment of equivalence follows. Judgmental designs are a preliminary approach. Additional checks, such as DIF or other statistical analyses, are also needed (Reynolds, Lowe et al., 1999).

Three statistical designs are available, depending on the characteristics of the sample. In the bilingual examinees design, participants who take both the original and the target version of the test are bilingual (Hambleton & Bollwark, 1991). In the source and target language monolinguals design, monolinguals in the original language take the original or back-translated version, and monolinguals in the target language take the target version (Ellis, 1991). In the third design, monolinguals in the original language take the original and back-translated versions.

After administration and scoring, statistical procedures are selected and performed to assess DIF. Procedures can include factor analysis, item response theory, logistic regression, and the Mantel-Haenszel technique. If DIF is statistically significant, additional analyses are necessary to investigate possible bias or lack of equivalence for different cultures or languages.

A study by Arnold, Montgomery, Castaneda, and Longoria (1994) illustrated the need to evaluate item equivalence. The researchers found that acculturation affected several subtests of the Halstead-Reitan neuropsychological test when used with unimpaired Hispanics. By contrast, Boivin et al. (1996) conducted a study with Lao children and identified variables such as nutritional development, parental education, and home environment that may influence scores on several tests, including the K-ABC, the Tactual Performance Test (TPT), and the computerized Tests of Variables of Attention (TOVA). These results suggested that tests can potentially be adapted to different cultures, although the challenges of doing so are formidable. Such results also showed that psychologists have addressed cultural equivalence issues for some time, contrary to the view of Helms (1992).

Hambleton and Zenisky (2011) offer a Review Form to evaluate the test translation and adaptation efforts. The form is comprised of 25 questions centering around five topics: General Translation Questions, Item Format and Appearance, Grammar and Phrasing, Passages and Other Item-Relevant Stimulus Materials (if relevant), and Cultural Relevance or Specificity. The Review Form reflects the complexity of the translation and adaptation process. For example, they cite research indicating that different font styles and typefaces can be a source of DIF.

## NATURE AND NURTURE

Part of the emotion surrounding the test bias controversy stems from its association in the human mind with the troubling notion of innate genetic inferiority. Given real differences, however, a genetic explanation is by no means inevitable. Absence of bias opens up the possibility of environmental causes as well, and explanations span the

sociopolitical spectrum. Discrimination, economic disadvantage, exclusion from educational opportunity, personal development, social support, practical information, and achievement-oriented values—all become possible causes, if differences are real.

All sides of the nature–nurture debate depend on the existence of real differences. Therefore, the debate will prove irresolvable unless the test bias question is somehow answered. The reverse, however, is not true. Test bias research can continue indefinitely with the nature–nurture question unresolved. Psychometricians are attempting to disentangle the nature–nurture debate from the empirical investigation of test bias, but the separation is unlikely to be a neat one (Reynolds, Lowe et al., 1999).

## CONCLUSIONS AND RECOMMENDATIONS

The conclusion reached in most of the research reviewed above was that test bias did not exist. Today, the same research would lead to different conclusions. Test bias exists but is small, which raises questions about its importance. It most often overestimates or over predicts minority examinees' performance, so that its social consequences may be very different from those typically ascribed to it, and appropriate responses to it may differ from those typically made. Finally, just as purely genetic and environmental paradigms have given way, the interpretation of zero bias should cede to a better informed understanding that bias cannot be understood in isolation from other possible influences.

We recommend that rigorous examination of possible test bias and inaccuracy should continue, employing the latest and most diverse techniques. Nonetheless, we caution against labeling tests biased in the absence of, or in opposition to, reliable evidence. To do so is of questionable effectiveness in the struggle to identify and combat real discrimination and to ensure that everyone is treated fairly.

Discrimination is a legitimate and compelling concern. We do not argue that it is rare, unimportant, or remotely acceptable. We do, however, suggest from research findings that standardized test bias is not a major source of discrimination. Accordingly, resources meant to identify and alleviate discrimination might better be directed toward real-world causes rather than standardized tests. In addition, we question whether the goal of equal opportunity is served if possible evidence of discrimination, or of inequalities resulting from it, is erased by well-meaning test publishers or other professionals.

The issue of bias in mental testing, too, is an important concern with strong historical precedence in the social sciences and with formidable social consequences. The controversy is liable to persist as long as we entangle it with the nature–nurture question and stress mean differences in standardized test scores. Similarly, the use of aptitude and achievement measures is long-standing and widespread, extending back more than 2,000 years in some cultures and across most cultures today. It is unlikely to disappear soon.

The news media may be partly responsible for a popular perception that tests and testing are uniformly biased or unfair. As indicated by the findings reviewed here, the view that tests are substantially biased has little support at present, at least in cultures with a common language and a degree of common experience. In addition, public pressure has pushed the scientific community to refine its definitions of bias, scrutinize the practices used to minimize bias in tests, and develop increasingly sophisticated statistical techniques to detect bias (Reynolds, Lowe et al., 1999; Samuda, 1975). Finally, the findings reviewed here give indications that fair testing is an attainable goal, albeit a challenging one that demands skill and training.

Reynolds, Lowe et al. (1999) suggested four guidelines to help ensure equitable assessment:

1. Investigate possible referral source bias, because evidence suggests that people are not always referred for services on impartial, objective grounds.
2. Inspect test developers' data for evidence that sound statistical analyses for bias have been completed.
3. Conduct assessments with the most reliable measure available.
4. Assess multiple abilities and use multiple methods.

In summary, clinicians should use accurately derived data from multiple sources before making decisions about an individual.

Clinicians should be cognizant of a person's environmental background and circumstances. Information about a client's home, community, and the like must be evaluated in an individualized decision-making process. Likewise, clinicians should not ignore evidence that disadvantaged, ethnic minority clients with unfavorable test results are as likely to encounter difficulties as are middle-class, majority clients with unfavorable test results, given the same environmental circumstances. The purpose of the assessment process is to beat the prediction—to suggest hypotheses for interventions that will prevent a predicted failure or adverse outcome (Reynolds, Lowe et al., 1999). This perspective,

although developed primarily around ability testing, is relevant to personality testing as well.

We urge clinicians to use tests fairly and in the interest of examinees, but we see little benefit in discarding standardized tests entirely. We recommend that test consumers evaluate each measure separately to ensure that results pertaining to bias are available and satisfactory. If results are unsatisfactory, local norming may produce less biased scores. If results are unavailable, additional testing may be possible, given samples of sufficient size. In addition, clinical practice and especially research should reflect an understanding of the conceptual distinctions, such as bias versus unfairness, described in this chapter.

A philosophical perspective emerging in the bias literature is that, before publication, test developers should not only demonstrate content, construct, and predictive validity but should also conduct content analysis in some form to ensure that offensive material is absent from the test. Expert reviews of test content can have a role, and the synergistic relationship between test use and psychometrics must be accommodated in an orderly manner before tests gain increased acceptance in society.

Nevertheless, informal reviews cannot meet the need to assess for bias. Test authors and publishers must demonstrate factorial congruence with all groups for whom a test is designed, to permit accurate interpretation. Comparisons of predictive validity with ethnic and gender groups are also important. Such research should take place during test development, a window during which measures can be altered using numerous item analysis procedures to minimize gender or ethnic bias. This practice has been uncommon, except with some recent achievement tests.

Greater attention to bias issues and personality tests are needed, though studies have emerged in recent years (e.g., Reynolds & Kamphaus, 1992; Suzuki & Ponterotto, 2008). Increased research is needed also for neuropsychological tests, for ability and achievement tests not yet investigated, for SES, and for minority examinees tested by majority examiners. Future results, it is expected, will continue to indicate consistency for different genders, races, ethnicities, and similar groups.

Finally, a clear consensus on fairness, and on steps to be taken to attain it, is needed between persons with humanitarian aims and those with scientific interest in test bias. Accommodation toward this end would ensure that everyone concerned with a given test was satisfied that it was unbiased and that the steps taken to achieve fairness could be held up to public scrutiny without reservation (Reynolds, Lowe et al., 1999). Test bias and fairness is a domain in great need of consensus, and this goal is attainable only with concessions on all sides.

## REFERENCES


Adebimpe, V. R., Gigandet, J., & Harris, E. (1979). MMPI diagnosis of black psychiatric patients. *American Journal of Psychiatry*, *136*, 85–87.

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

American Psychological Association. (2002). *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists*. Retrieved from www.apa.org/pi/oema/resources/policy/multicultural-guidelines.aspx

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York, NY: Macmillan.

Arnold, B., Montgomery, G., Castaneda, I., & Longoria, R. (1994). Acculturation and performance of Hispanics on selected Halstead-Reitan neuropsychological tests. *Assessment*, *1*, 239–248.

Binet, A., & Simon, T. (1973). *The development of intelligence in children*. New York, NY: Arno. (Original work published 1916)

Boivin, M., Chounramany, C., Giordani, B., Xaisida, S., Choulamountry, L., Pholsena, P., et al. (1996). Validating a cognitive ability testing protocol with Lao children for community development applications. *Neuropsychology*, *10*, 1–12.

Bossard, M., Reynolds, C. R., & Gutkin, T. B. (1980). A regression analysis of test bias on the Stanford-Binet Intelligence Scale. *Journal of Clinical Child Psychology*, *9*, 52–54.

Bouchard, T. J., & Segal, N. L. (1985). *Environment and IQ*. In B. Wolman (Ed.), *Handbook of intelligence* (pp. 391–464). New York, NY: Wiley-Interscience.

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.

Brooks, A. P. (1997). TAAS unfair to minorities, lawsuit claims. *Austin American-Statesman*, p. A1.

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.

Butcher, J. N. (1982). Cross-cultural research methods in clinical psychology. In P. C. Kendall & J. N. Butcher (Eds.), *Black children: Social educational and parental environments* (pp. 33–51). Beverly Hills, CA: Sage.

Butcher, J. N. (2009). Clinical personality assessment: History, evolution, contemporary models, and practical applications. In. J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 5–21). New York, NY: Oxford University Press.

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

Cattell, R. B. (1979). Are culture fair intelligence tests possible and necessary? *Journal of Research and Development in Education*, *12*, 3–13.

Cheung, F. M., Leong, F. T. L., & Ben-Porath, Y. S. (2003). Psychological assessment in Asia: Introduction to the special section. *Psychological Assessment*, *15*, 243–247.

Chipman, S., Marshall, S., & Scott, P. (1991). Content effect on word-problem performance: A possible source of test bias? *American Educational Research Journal*, *28*, 897–915.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated universities. *Journal of Educational Measurement*, *5*, 118–124.

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

*Cortez v. Rosen* (N.D. Cal., Docket No. C-73-388-SW, March 11, 1975).

Cronbach, L. J., & Drenth, P. J. D. (Eds.). (1972). *Mental tests and cultural adaptation*. The Hague, the Netherlands: Mouton.

Dana, R. H. (2005). *Multicultural assessment: Principles, application, and examples*. Mahwah, NJ: Erlbaum.

Dean, R. S. (1979, September). *WISC-R factor structure for Anglo and Hispanic children*. Paper presented at the annual meeting of the American Psychological Association, New York.

DeFries, J. C., Vandenberg, S. G., McClearn, G. E., Kuse, A. R., Wilson, J. R., Ashton, G. C., et al. (1974). Near identity of cognitive structure in two ethnic groups. *Science*, *183*, 338–339.

Dent, H. E. (1996). Non-biased assessment or realistic assessment? In R. L. Jones (Ed.), *Handbook of tests and measurement for Black populations* (Vol. 1, pp. 103–122). Hampton, VA: Cobb & Henry.

Ebel, R. L. (1979). Intelligence: A skeptical view. *Journal of Research and Development in Education*, *12*, 14–21.

Educational Testing Service. (2002). *Standards for quality and fairness and international principles for fairness review of assessments*. Retrieved from www.ets.org/Media/About_ETS/pdf/frintl.pdf

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). *Intelligence and cultural differences: A study of cultural learning and problem-solving*. Chicago, IL: University of Chicago Press.

Edwards, O. W., & Oakland, T. D. (2006). Factorial invariance of Woodcock-Johnson III scores for African Americans and Caucasian Americans. *Journal of Psychoeducational Assessment*, *24*(4), 358–366.

Elliot, R. (1987). *Litigating intelligence*. Dover, MA: Auburn House.

Ellis, B. B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. *Bulletin of the International Test Commission*, *18*, 33–51.

Figueroa, R. A. (1991). Bilingualism and psychometrics. *Diagnostique*, *17*(1), 70–85.

Fine, B. (1975). *The stranglehold of the IQ*. Garden City, NY: Doubleday.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). San Francisco, CA: Wiley.

Flynn, J. R. (1991). *Asian-Americans: Achievement beyond IQ*. Hillsdale, NJ: Erlbaum.

Freedle, R. O. (2003). Correcting the SATs ethnic and social class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73*(1), 1–43.

Freedle, R. O. (2010). On replicating ethnic test bias effects: The Santelices and Wilson study. *Harvard Educational Review*, *80*(3), 394–403.

Freedle, R. O., & Kostin, I. (1997). Predicting Black and White differential item functioining in verbal analogy performance. *Intelligence*, *24*(3), 417–444.

Geisinger, K. F. (2005). The testing industry, ethnic minorities, and individuals with disabilities. In R. P. Phelps (Ed.) *Defending standardized testing* (pp. 187–204). Mahwah, NJ: Erlbaum.

Gerkin, K. C. (1978). Performance of Mexican-American children on intelligence tests. *Exceptional Children*, *44*, 438–443.

Gopaul-McNicol, S. & Armour-Thomas, E. (2002). *Assessment and culture: Psychological tests with minority populations*. New York, NY: Academic Press.

Gould, S. J. (1981). *The mismeasure of man*. New York, NY: Norton.

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.

Graves, S., & Mitchell, A. (2011). Is the moratorium over? African American psychology professionals' views on intelligence testing in response to changes in federal policy. *Journal of Black Psychology*, *37*(4), 407–425.

Gross, M. (1967). *Learning readiness in two Jewish groups*. New York, NY: Center for Urban Education.

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Guterman, S. S. (1979). IQ tests in research on social stratification: The cross-class validity of the tests as measures of scholastic aptitude. *Sociology of Education*, *52*, 163–173.

Gutkin, T. B., & Reynolds, C. R. (1980). Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services. *Journal of School Psychology*, *18*, 34–39.

Gutkin, T. B., & Reynolds, C. R. (1981). Factorial similarity of the WISC-R for white and black children from the standardization sample. *Journal of Educational Psychology*, *73*, 227–231.

Hagie, M. U., Gallipo, P. L., & Svien, L. (2003). Traditional culture versus traditional assessment for American Indian students: An investigation of potential test item bias. *Assessment for Effective Intervention*, *29*(1), 15–25.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment*, *9*, 54–65.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*, 229–244.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Test Commission*, *18*, 3–32.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for adaptations. *European Journal of Psychological Assessment*, *11*, 147–157.

Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46–74). New York, NY: Cambridge University Press.

Hammill, D. D. (1991). *Detroit Tests of Learning Aptitude* (3rd ed.). Austin, TX: Pro-Ed.

Hammill, D. D., Pearson, N. A., & Wiederholt, J. L. (1997). *Comprehensive test of nonverbal intelligence*. Austin, TX: Pro-Ed.

Harrington, G. M. (1968a). Genetic-environmental interaction in "intelligence": I. Biometric genetic analysis of maze performance of *Rattus norvegicus*. *Developmental Psychology*, *1*, 211–218.

Harrington, G. M. (1968b). Genetic-environmental interaction in "intelligence": II. Models of behavior, components of variance, and research strategy. *Developmental Psychology*, *1*, 245–253.

Harrington, G. M. (1975). Intelligence tests may favor the majority groups in a population. *Nature*, *258*, 708–709.

Harrington, G. M. (1976, September). *Minority test bias as a psychometric artifact: The experimental evidence*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Hartlage, L. C., Lucas, T., & Godwin, A. (1976). Culturally biased and culturally fair tests correlated with school performance in culturally disadvantaged children. *Journal of Consulting and Clinical Psychology*, *32*, 325–327.

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

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial/group assessment: A quantitative perspective. *American Psychologist*, *61*, 845–859.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. New York, NY: Free Press.

Hilliard, A. G. III. (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.

Hilliard, A. G. III. (1984). IQ testing as the emperor's new clothes: A critique of Jensen's Bias in Mental Testing. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 139–169). New York, NY: Plenum Press.

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.

Jackson, G. D. (1975). Another psychological view from the Association of Black Psychologists. *American Psychologist*, *30*, 88–93.

James, B. J. (1995). *A test of Harrington's experimental model of ethnic bias in testing applied to a measure of emotional functioning in adolescents*. (Unpublished doctoral dissertation). Texas A&M University, College Station.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, *39*, 1–123.

Jensen, A. R. (1974). How biased are culture loaded tests? *Genetic Psychology Monographs*, *90*, 185–224.

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

Jensen, A. R. (1977). An examination of culture bias in the Wonderlic Personnel Test. *Intelligence*, *1*, 51–64.

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

Jensen, A. R. (1984). Test bias: Concepts and criticisms. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 507–586). New York, NY: Plenum Press.

Kaiser, S. (1986). *Ability patterns of black and white adults on the WAIS-R independent of general intelligence and as a function of socioeconomic status*. (Unpublished doctoral dissertation). Texas A&M University, College Station.

Kamphaus, R. W., & Reynolds, C. R. (1987). *Clinical and research applications of the K-ABC*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S. (1973). Comparison of the performance of matched groups of black children and white children on the Wechsler Preschool and Primary Scale of Intelligence. *Journal of Consulting and Clinical Psychology*, *41*, 186–191.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York, NY: Wiley-Interscience.

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

Kaufman, A. S., & Kaufman, N. L. (1973). Black-white differences on the McCarthy Scales of Children's Abilities. *Journal of School Psychology*, *11*, 196–206.

Keith, T. Z., & Reynolds, C. R. (1990). Measurement and design issues in child assessment research. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children*. New York, NY: Guilford Press.

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. (1975). *Race differences in intelligence*. San Francisco, CA: Freeman.

Lonner, W. J. (1985). Issues in testing and assessment in crosscultural counseling. *Counseling Psychologist*, *13*, 599–614.

Lynn, R. (1995). Cross-cultural differences in intelligence and personality. In D. Sakolfske & M. Zeidner (Eds.), *The international handbook of personality and intelligence* (pp. 107–134). New York, NY: Plenum Press.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*, 95–107.

Mainstream science on intelligence. (1994, December 13). *Wall Street Journal*, p. A18.

Maller, S. J. (2003). Best practices in detecting bias in nonverbal tests. In R. S. McCallum (Ed.) *Handbook of nonverbal assessment* (pp. 23–48). New York, NY: Plenum Press.

Marjoribanks, K. (1972). Ethnic and environmental influences on mental abilities. *American Journal of Sociology*, *78*, 323–337.

Mayfield, J. W., & Reynolds, C. R. (1998). Are ethnic differences in diagnosis of childhood psychopathology an artifact of psychometric methods? An experimental evaluation of Harrington's hypothesis using parent report symptomatology. *Journal of School Psychology*, *36*, 313–334.

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc Cross Battery Assessment*. Boston, MA: Allyn & Bacon.

McGurk, F. V. J. (1951). *Comparison of the performance of Negro and white high school seniors on cultural and noncultural psychological test questions*. Washington, DC: Catholic University of American Press.

McShane, D. (1980). A review of scores of American Indian children on the Wechsler Intelligence Scale. *White Cloud Journal*, *2*, 18–22.

Mercer, J. R. (1976, August). *Cultural diversity, mental retardation, and assessment: The case for nonlabeling*. Paper presented at the Fourth International Congress of the International Association for the Scientific Study of Mental Retardation, Washington, DC.

Mercer, J. R. (1979). *System of Multicultural Pluralistic Assessment (SOMPA): Conceptual and technical manual*. San Antonio, TX: Psychological Corporation.

Miele, F. (1979). Cultural bias in the WISC. *Intelligence*, *3*, 149–164.

Monnot, M. J., Quirk, S. W., Hoerger, M. & Brewer, L. (2009). Racial bias in personality assessment: Using the MMPI-2 to predict psychiatric diagnoses of African American and Caucasian chemical dependency inpatients. *Psychological Assessment*, *21*(2), 137–151.

Morales, E. S., & George, C. (1976, September). *Examiner effects in the testing of Mexican-American children*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Moran, M. P. (1990). The problem of cultural bias in personality assessment. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children. Vol. 2: Personality, behavior, and context* (pp. 524–545). New York, NY: Guilford Press.

Mpofu, E., & Ortiz, S. O. (2009). Equitable assessment practices in diverse contexts. In E. L. Grigorenko (Ed.), *Multicultural psychoeducational assessment* (pp. 41–76). New York, NY: Springer.

Nandakumar, R., Glutting, J. J., & Oakland, T. (1993). Mantel-Haenszel methodology for detecting item bias: An introduction and example using the Guide to the Assessment of Test Session Behavior. *Journal of Psychoeducational Assessment*, *11*(2), 108–119.

Neisser, U., Boodoo, G., Bouchard, T. J. Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (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.

Oakland, T., & Feigenbaum, D. (1979). Multiple sources of test bias on the WISC-R and the Bender-Gestalt Test. *Journal of Consulting and Clinical Psychology*, *47*, 968–974.

Oakland, T., & Matuszek, P. (1977). Using tests in nondiscriminatory assessment. In T. Oakland (Ed.), *Psychological and educational assessment of minority children*. New York, NY: Brunner/Mazel.

Okazaki, S., & Sue, S. (2000). Implications for test revisions for assessment with Asian Americans. *Psychological Assessment*, *12*(30), 272–280.

Ortiz, S. O., & Ochoa, S. H. (2005). Advances in cognitive assessment of culturally linguistically diverse individuals. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories,*

*tests and issues* (2nd ed., pp. 234–250). New York, NY: Guilford Press.

Padilla, A. M. (1988). Early psychological assessment of Mexican-American children. *Journal of the History of the Behavioral Sciences*, *24*, 113–115.

Payne, B., & Payne, D. (1991). The ability of teachers to identify academically at-risk elementary students. *Journal of Research in Childhood education*, *5*(2), 116–126.

Pintner, R. (1931). *Intelligence testing*. New York, NY: Holt, Rinehart, & Winston.

Poortinga, Y. H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S. H. Irvine & J. W. Berry (Eds.), *Human assessment and cross-cultural factors* (pp. 237–258). New York, NY: Plenum Press.

Poortinga, Y. H., & Malpass, R. S. (1986). Making inferences from cross-cultural data. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 17–46). Beverly Hills, CA: Sage.

Prifitera, A., & Saklofske, D. H. (Eds.). (1998). *WISC-III clinical use and interpretation: Scientist-practitioner perspectives*. San Diego, CA: Academic Press.

Prifitera, A., Weiss, L. G., & Saklofske, D. H. (1998). The WISC-III in context. In A. Prifitera & D. H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 1–37). San Diego, CA: Academic Press.

Ramsay, M. C. (1997, November). *Structural equation modeling and test bias*. Paper presented at the annual meeting of the Educational Research Exchange, Texas A&M University, College Station.

Ramsay, M. C. (1998a, February). *Proposed theories of causation drawn from social and physical science epistemology*. Paper presented at the annual meeting of the Education Research Exchange, Texas A&M University, College Station.

Ramsay, M. C. (1998b, February). *The processing system in humans: A theory*. Paper presented at the annual meeting of the Education Research Exchange, Texas A&M University, College Station.

Ramsay, M. C. (2000). *The putative effects of smoking by pregnant women on birthweight, IQ, and developmental disabilities in their infants: A methodological review and multivariate analysis*. (Unpublished doctoral dissertation). Texas A&M University, College Station.

Ramsay, M. C., & Reynolds, C. R. (1995). Separate digits tests: A brief history, a literature review, and a reexamination of the factor structure of the Test of Memory and Learning (TOMAL). *Neuropsychology Review*, *5*, 151–171.

Ramsay, M. C., & Reynolds, C. R. (2000a). Development of a scientific test: A practical guide. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (3rd ed.). Amsterdam, the Netherlands: Pergamon Press.

Ramsay, M. C., & Reynolds, C. R. (2000b). Does smoking by pregnant women influence birth weight, IQ, and developmental disabilities in their infants? A methodological review and multivariate analysis. *Neuropsychology Review*, *10*, 1–49.

Ramsden, S., Richardson, F. M., Josse, G., Thomas, M. S., Ellis, C., Shakeshaft, C., et al. (2011). Verbal and nonverbal changes to the teenage brain. *Nature*, *479*, 113–116.

Reschly, D. J. (1978). WISC-R factor structures among Anglos, Blacks, Chicanos, and Native American Papagos. *Journal of Consulting and Clinical Psychology*, *46*, 417–422.

Reschly, D. J. (1997). Diagnostic and treatment utility of intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 437–456). New York, NY: Guilford Press.

Reschly, D. J. (2000). PASE v. Hannon. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Encyclopedia of special education* (2nd ed., pp. 1325–1326). New York, NY: Wiley.

Reschly, D. J., & Sabers, D. (1979). Analysis of test bias in four groups with the regression definition. *Journal of Educational Measurement*, *16*, 1–9.

Reynolds, C. R. (1978). *Differential validity of several preschool assessment instruments for blacks, whites, males, and females*. (Unpublished doctoral dissertation). University of Georgia, Athens.

Reynolds, C. R. (1980). Differential construct validity of intelligence as popularly measured: Correlation of age and raw scores on the WISC-R for blacks, whites, males and females. *Intelligence: A Multidisciplinary Journal*, *4*, 371–379.

Reynolds, C. R. (1982a). Construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 199–227). Baltimore, MD: Johns Hopkins University Press.

Reynolds, C. R. (1982b). 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. (1995). Test bias in the assessment of intelligence and personality. In D. Saklofsky & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 545–576). New York, NY: Plenum Press.

Reynolds, C. R. (1998a). Cultural bias in testing of intelligence and personality. In A. Bellack & M. Hersen (Series Eds.) & C. Belar (Vol. Ed.), *Comprehensive clinical psychology: Vol. 10. Cross cultural psychology* (pp. 53–92). New York, NY: Elsevier Science.

Reynolds, C. R. (1998b). Need we measure anxiety separately for males and females? *Journal of Personality Assessment*, *70*, 212–221.

Reynolds, C. R. (1999). Cultural bias in testing of intelligence and personality. In M. Hersen & A. Bellack (Series Eds.) & C. Belar (Vol. Ed.), *Comprehensive clinical psychology: Vol. 10. Sociocultural and individual differences* (pp. 53–92). Oxford, UK: Elsevier Science.

Reynolds, C. R. (2000a). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology* (pp. 249–285). New York, NY: Kluwer Academic/Plenum Press.

Reynolds, C. R. (2000b). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law*, *6*, 144–150.

Reynolds, C. R. (2001). *Professional manual for the Clinical Assessment Scales for the Elderly*. Odessa, FL: Psychological Assessment Resources.

Reynolds, C. R., & Bigler, E. D. (1994). *Test of Memory and Learning (TOMAL)*. Austin, TX: Pro-Ed.

Reynolds, C. R., & Brown, R. T. (1984a). Bias in mental testing: An introduction to the issues. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 1–39). New York, NY: Plenum Press.

Reynolds, C. R., & Brown, R. T. (1984b). *Perspectives on bias in mental testing*. New York, NY: Plenum Press.

Reynolds, C. R., & Carson, A. D. (2005). Methods for assessing cultural bias in tests. In C. L. Frisby & C. R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 795–839). Hoboken, NJ: Wiley.

Reynolds, C. R., Chastain, R., Kaufman, A. S., & McLean, J. (1987). Demographic influences on adult intelligence at ages 16 to 74 years. *Journal of School Psychology*, *25*, 323–342.

Reynolds, C. R., & Gutkin, T. B. (1980, September). *WISC-R performance of blacks and whites matched on four demographic variables*. Paper presented at the annual meeting of the American Psychological Association, Montreal, Canada.

Reynolds, C. R., & Gutkin, T. B. (1981). A multivariate comparison of the intellectual performance of blacks and whites matched on four demographic variables. *Personality and Individual Differences*, *2*, 175–180.

Reynolds, C. R., & Harding, R. E. (1983). Outcome in two large sample studies of factorial similarity under six methods of comparison. *Educational and Psychological Measurement*, *43*, 723–728.

Reynolds, C. R., & Hartlage, L. C. (1979). Comparison of WISC and WISC-R regression lines for academic prediction with black and white referred children. *Journal of Consulting and Clinical Psychology*, *47*, 589–591.

Reynolds, C. R., & Jensen, A. R. (1983, September). *Patterns of intellectual performance among blacks and whites matched on* "g." Paper presented at the annual meeting of the American Psychological Association, Montreal, Canada.

Reynolds, C. R., & Kaiser, S. (1992). Test bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed., pp. 487–525). New York, NY: Wiley.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children (BASC): Manual*. Circle Pines, MN: American Guidance Service.

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., 549–595). New York, NY: Wiley.

Reynolds, C. R., & Paget, K. (1981). Factor analysis of the Revised Children's Manifest Anxiety Scale for blacks, whites, males, and females with a national normative sample. *Journal of Consulting and Clinical Psychology*, *49*, 349–352.

Reynolds, C. R., Willson, V. L., & Chatman, S. P. (1984). Item bias on the 1981 revisions of the Peabody Picture Vocabulary Test using a new method of detecting bias. *Journal of Psychoeducational Assessment*, *2*, 219–221.

Reynolds, C. R., Willson, V. L., & Chatman, S. P. (1985). Regression analyses of bias on the Kaufman Assessment Battery for Children. *Journal of School Psychology*, *23*, 195–204.

Reynolds, C. R., Willson, V. L., & Ramsay, M. C. (1999). Intellectual differences among Mexican Americans, Papagos and Whites, independent of g. *Personality and Individual Differences*, *27*, 1181–1187.

Richardson, T. Q. (1995). The window dressing behind the bell curve. *School Psychology Review*, *24*, 42–44.

Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale-Revised: Examiner's manual*. Wood Dale, IL: Stoelting.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*(2), 235–294.

Samuda, A. J. (1975). *Psychological testing of American minorities: Issues and consequences*. New York, NY: Dodd.

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review*, *80*(1), 106–128.

Sandoval, J. (1979). The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology*, *47*, 919–927.

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.

Sattler, J. M. (1974). *Assessment of children's intelligence*. Philadelphia, PA: Saunders.

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Author.

Sattler, J. M., & Gwynne, J. (1982). White examiners generally do not impede the intelligence test performance of black children: To debunk a myth. *Journal of Consulting and Clinical Psychology*, *50*, 196–208.

Scarr, S. (1981). Implicit messages: A review of bias in mental testing. *American Journal of Education*, *89*(3), 330–338.

Scheuneman, J. D. (1987). An experimental, exploratory study of the causes of bias in test items. *Journal of Educational Measurement*, *29*, 97–118.

Schmidt, W. H. (1983). Content biases in achievement tests. *Journal of Educational Measurement*, *20*, 165–178.

Schoenfeld, W. N. (1974). Notes on a bit of psychological nonsense: "Race differences in intelligence." *Psychological Record*, *24*, 17–32.

Scholwinski, E. (1985). *Ability patterns of blacks and whites as determined by the subscales on the Wechsler Adult Intelligence Scale–Revised*. (Unpublished doctoral dissertation). Texas A&M University, College Station.

Shuey, A. M. (1966). *The testing of Negro intelligence* (2nd ed.). New York, NY: Social Science Press.

Spitz, H. (1986). *The raising of intelligence*. Hillsdale, NJ: Erlbaum.

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., & Aronson, J. (2004). Stereotype threat does not live by Steele and Aronson alone. *American Psychologist*, *59*(1), 47–48.

Stern, W. (1914). *The psychological methods of testing intelligence*. Baltimore, MD: Warwick & York.

Sternberg, R. J. (1980). Intelligence and test bias: Art and science. *Behavioral and Brain Sciences*, *3*, 353–354.

Suzuki, L. A., & Aronson, J. (2005). The cultural malleability of intelligence and its impact on the racial/ethnic hierarchy. *Psychology, Public Policy, and Law*, *11*(2), 320–327.

Suzuki, L. A., Kugler, J. F., & Aguiar, L. J. (2005). Assessment practices in racial-cultural psychology. In R. T. Carter (Ed.) *Handbook of racial-cultural psychology and counseling: Training and practice* (Vol. 2, pp. 297–315). Hoboken, NJ: Wiley.

Suzuki, L. A., Onoue, M. A., & Hill, J. (forthcoming). Clinical assessment: A multicultural perspective. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*. Washington, DC: APA Books.

Suzuki, L. A., & Ponterotto, J. G. (Eds.) (2008). *Handbook of multicultural assessment: Clinical, psychological and educational applications* (3rd ed.). San Francisco, CA: Sage.

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

Torrance, E. P. (1980). Psychology of gifted children and youth. In W. M. Cruickshank (Ed.), *Psychology of exceptional children and youth*. Englewood Cliffs, NJ: Prentice-Hall.

Tyler, L. E. (1965). *The psychology of human differences*. New York, NY: Appleton-Century-Crofts.

Valencia, R. A., & Suzuki, L. A. (2001). *Intelligence testing with minority students: Foundations, performance factors and assessment issues*. Thousand Oaks, CA: Sage.

Valencia, R. A., Suzuki, L. A., & Salinas, M. (2001). Test bias. In R. R. Valencia & L. A. Suzuki, *Intelligence testing with minority students: Foundations, performance factors and assessment issues* (pp. 111–181). Thousand Oaks, CA: Sage.

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologists*, *1*, 89–99.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Culture-free measurement in the history of cross-cultural psychology. *Bulletin of the International Test Commission*, *18*, 72–87.

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto and F. J. R. van de Vijver (Eds.) *Cross-cultural research*

*methods in psychology* (pp. 17–45). New York, NY: Cambridge University Press.

Verney, S. P., Granholm, E., Marshall, S. P., Malcarne, V. L., & Saccuzzo, D. P. (2005). Culture-fair cognitive ability assessment: Information processing and psychophysiological approaches. *Assessment*, *12*, 303–319.

Wechsler, D. (1975). Intelligence defined and undefined: A relativistic appraisal. *American Psychologist*, *30*, 135–139.

Weiss, L. G., & Prifitera, A. (1995). An evaluation of differential prediction of WIAT achievement scores from WISC-III FSIQ across ethnic and gender groups. *Journal of School Psychology*, *33*, 297–304.

Weiss, L. G., Prifitera, A., & Roid, G. H. (1993). The WISC-III and fairness of predicting achievement across ethnic and gender groups. *Journal of Psychoeducational Assessment*, 35–42.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*(5), 696–719.

Wigdor, A. K., & Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies*. Washington, DC: National Academy of Sciences.

Willerman, L. (1979). *The psychology of individual and group differences*. San Francisco, CA: Freeman.

Williams, R. L. (1970). Danger: Testing and dehumanizing Black children. *Clinical Child Psychology Newsletter*, *9*, 5–6.

Williams, R. L. (1974). From dehumanization to black intellectual genocide: A rejoinder. In G. J. Williams & S. Gordon (Eds.), *Clinical child psychology: Current practices and future perspectives*. New York, NY: Behavioral Publications.

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.

Willson, V. L., Nolan, R. F., Reynolds, C. R., & Kamphaus, R. W. (1989). Race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of School Psychology*, *27*, 289–296.

Wright, B. J., & Isenstein, V. R. (1977–1978). *Psychological tests and minorities* (DHEW Publication No. ADM 78–482). Rockville, MD: National Institutes of Mental Health.

Zieky, M. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 359–376). Mahwah, NJ: Erlbaum.