School Psychology Quarterly, Vol. 14, No. 3, 1999, pp. 208–238

Bias in Mental Testing since *Bias in Mental Testing*

Robert T. Brown
University of North Carolina at Wilmington

Cecil R. Reynolds and Jean S. Whitaker
Texas A&M University

Racial/ethnic subgroup differences in average performance on standardized tests of cognitive ability are well established (Gordon & Bhattacharya, "Race and Intelligence," and Jensen, "Race and IQ Scores," in *Encyclopedia of Human Intelligence, 1994; Herrnstein & Murray, The Bell Curve,* 1994), but the reasons for these differences are an ongoing source of controversy. One popular and longstanding claim is that mean differences are caused by "cultural bias" in the tests. Arthur Jensen exhaustively reviewed the empirical literature on the issue of test bias, which resulted in his seminal book, *Bias in Mental Testing* (BIMT), published in 1980. On the basis of empirical criteria for evaluating test bias, Jensen concluded that standardized aptitude/ability tests predict equally well for American-born, English-speaking majority and minority subgroups and measure similar constructs. This paper summarizes the major conclusions from BIMT and evaluates writing on test bias published since BIMT. We conclude that empirical research to date consistently finds that standardized cognitive tests are not biased in terms of predictive and construct validity. Furthermore, continued claims of test bias, which appear in academic journals, the popular media, and some psychology textbooks, are not empirically justified. These claims of bias should be met with skepticism and evaluated critically according to established scientific principles.

> There is perhaps no other assessment issue as heated, controversial, and frequently debated as that of bias in cognitive assessment. (Taylor, 1991, p. 3)

The interpretation of mean differences in mental test scores for native-born, English-speaking ethnic/racial subgroups is a source of major professional and public controversy (e.g., Jacoby & Glauberman, 1995). According to the cultural test bias

The authors thank Len Lecci and an anonymous reviewer for helpful comments on an earlier draft, and Craig Frisby for his thoughtful discussions, suggestions, and great tolerance during development of this article. The authors of course accept responsibility for all content.

Address correspondence to either Robert T. Brown, Department of Psychology, UNCW, Wilmington, NC 28403-3297 (brown@uncwil.edu) or Cecil R. Reynolds, Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225 (crrh@bluebon.net).

hypothesis (CTBH), deficits in mean scores of these subgroups relative to the majority group reflect no real differences in ability, but rather problems in the construction, design, administration, or interpretation of tests. Thus, resulting scores are less valid for minority subgroups. Reynolds, Lowe, and Saenz (1999, pp. 556–557) divide claims of the CTBH into the following seven categories:

1. *Inappropriate content.* Since cognitive tests are designed for the cultural values and practices of middle-class White children, non-White and/or lower-class children will be at a disadvantage (and are more likely to perform poorly) because of a lack of exposure to test questions or test-related stimulus materials.
2. *Inappropriate standardization samples.* Persons from racial/ethnic minority groups may be underrepresented in standardization samples relative to their proportions in the overall population on whom the test will be used. Furthermore, even if persons from minority/ethnic groups are represented proportionally, their absolute numbers may be too small to prevent bias in item selection.
3. *Examiners' and language bias.* White examiners who speak standard English may not communicate effectively with minority examinees, which may cause examiners to penalize unfairly minority examinees in scoring or cause examinees to underperform for the examiner.
4. *Inequitable social consequences.* Biased tests will result in minority group members being relegated to "dead end" educational tracks and/or suffer from the effects of labeling.
5. *Measurement of different constructs.* When cognitive tests are used with individuals who are not part of the "majority" culture on which the tests are based, they are measuring a different construct than that intended by the test developer. That is, standardized mental tests do not measure intelligence in minority groups as they do in the majority group.
6. *Differential predictive validity.* Tests do not accurately predict relevant criteria for minority group members. Additionally, the criteria against which tests are typically correlated (for majority group members) are themselves biased against minority group members.
7. *Qualitatively distinct minority and majority aptitude and personality.* Cultural differences between racial/ethnic minorities and the majority are so profound as to require different conceptualizations of ability and personality.

OBJECTIONS TO STANDARDIZED TESTS PRIOR TO *BIAS IN MENTAL TESTING*

Responding to Jensen's (1969) controversial "How much can we boost IQ and scholastic achievement?" 18 well-known members of the Council of the Society for the Study of Social Issues (SPSSI, 1969) stated: "We must also recognize the

limitations of present day intelligence tests. Largely developed and standardized on white middle-class children, these tests tend to be biased against black children to an unknown degree" (p. 626). In 1968, the Association of Black Psychologists (ABP) sought a moratorium against the use of psychological and educational tests with African American children, and in 1969 adopted an official policy supporting African American parents who refused to allow their children or themselves to be assessed with such tests. In 1974, the NAACP adopted a resolution also demanding a moratorium, and the ABP's Committee on Testing issued a position paper urging, among other things: (a) cessation of testing with African Americans until culturally specific tests were available, and (b) removal of scores on standardized tests from records of African Americans.

Additional objections to standardized test usage with minority group and/or low SES children can be found in the pre-1980 literature (e.g., Hoffman, 1962; Lawler, 1978; Williams, 1971, 1974). These positions assume that much, if not all, of the mean differences in test scores between groups can be attributed to bias. A related assumption is that if minority members obtain low score distributions relative to Whites on a given test, then that test is *ipso facto* biased (e.g., Alley & Foster, 1978).

BASIC POSITIONS IN JENSEN'S *BIAS IN MENTAL TESTING*

Jensen (1984a) describes the state of test bias research prior to the 1970s:

> Prior to the 1970s, the treatment of test bias in the psychological literature was fragmentary, unsystematic, and conceptually confused. Clear and generally agreed-upon definitions of bias were lacking, as was a psychometrically defensible methodology for objectively recognizing test bias....The subject lacked the carefully thought-out rationale and statistical methodology that psychometrics had long invested in such topics as reliability, validity, and item selection. (p. 507)

Beginning in the 1970s, however, a large body of objective, empirical analyses of the complex issues involved in the use of standardized mental tests with American-born, English-speaking subgroups began to appear (e.g., see Berk, 1982).

Jensen's (1980) *Bias in Mental Testing* (BIMT) represented an exhaustive review of "empirical research relevant to the evaluation of cultural bias in psychological and educational tests that was available at the time that his book was prepared" (Reynolds & Brown, 1984, p. vii). Jensen (1984a) has stated that he wrote BIMT to accomplish three objectives: (a) establish clear and precise definitions of test bias; (b) explicate objective, operational psychometric criteria of bias and the statistical methods for detecting bias in tests; and (c) examine the results of applying bias detection methods to the then most widely used standardized tests in schools, colleges, armed services, and

civilian employment. In the next sections, we overview some of BIMT's more important conclusions related to racial/ethnic bias.

Orientation to the Topic of Test Bias

According to Jensen (1980), a proper approach to the scientific study of test bias can proceed only if common but fallacious assumptions are first identified and laid to rest. These common fallacies are (a) the egalitarian fallacy—all human subgroups are identical or equal in traits measured by tests; (b) the culture-bound fallacy—test items can be identified or graded as to their "culture-loadedness" from casual inspection and/or subjective judgment; and (c) the standardization fallacy—a test is necessarily biased when used with any population other than those included in large numbers in the standardization sample. Jensen (1980) further argued that test bias and fairness are separable issues: Fairness is a moral, legal, and/or philosophical issue on which reasonable persons legitimately disagree, whereas test bias is an *empirically based statistical* issue that deals with the psychometric properties of a given test as used with two or more specified subpopulations (p. 375).

Content Validity of Tests

Content-description validation procedures involve the systematic examination of the test content to determine whether the test item fits the behavior domain to be measured (e.g., Anastasi & Urbina, 1997). A common tactic of CTBH adherents is to claim, on rational analysis of individual test items, that some items are biased owing to their wording or content. One such item from the WISC-R comprehension subtest is the now notorious, "What is the thing to do if a boy/girl much smaller than yourself starts to fight with you?" The correct answer is "Walk away," or "Don't hit him back." The item was attacked as biased against inner-city African American children since they should hit the child back to maintain status. Perhaps the most extreme example of this approach was the development of putative "intelligence tests," such as the Black Intelligence Test of Cultural Homogeneity (BITCH) (Williams, 1974), based on language and experiences supposedly more common to African Americans than to Whites. The BITCH, however, has no apparent predictive or construct validity for any group, although it is still cited occasionally. Additionally, as Jensen reported, large-sample research indicates that of African American and White children of the same overall IQ, proportionately more African Americans give the correct answer to the "What is the thing to do if..." question indicating that the item frequently attacked as biased is actually relatively easier for African American children than for White children!

Predictive Validity of Tests

Predictive validity issues are the most important ones when dealing with the practical use of test scores in making decisions about selection for a particular educa-

tional setting or job. Jensen (1980) defined predictive validity bias as "*systematic error* (as contrasted to random errors of measurement) in the prediction of the criterion variable for persons of different subpopulations as a result of basing prediction on a common regression equation for all persons regardless of their subpopulation memberships..." (p. 380). If estimated true scores are used, a test will be a biased predictor if two or more groups differ significantly in slopes, intercepts, or standard error of estimates of the separate regression lines for those groups. When estimated true scores are not used, then the resulting unreliability of test scores (which is in all tests to some degree) will result in differences in regression intercepts (for groups with different mean performance on the predictor). This need not impede the evaluation of bias as long as separate regression lines are used for predicting the criterion.

After analyzing the available literature up to 1980, Jensen (1980) arrived at the following conclusions: (a) The large majority of studies did not find differential test validity for the two most researched groups, African Americans and Whites; (b) When significant differences in regression parameters were found between African Americans and Whites, the differences were in regression intercepts, with the African American intercept lower than the White one; and (c) This intercept bias results in *overprediction* of African Americans' criterion performance when predictions are based on a White or common regression line. This outcome would *favor* the African American group in a "colorblind" selection process, directly contrary to claims by CTBH advocates.

Construct Validity of Tests

Construct validity issues are relevant to a scientific understanding of what underlying psychological processes tests measure. Whereas investigations of predictive validity use data that are *external* to the tests (i.e., the criterion), investigations of construct validity evaluate *internal* psychometric indices. Examples of internal indices discussed in BIMT include item difficulty levels, item discrimination indices, item score × total score correlations, item characteristic curves (ICCs), reliability coefficients, and results of various factor analytic procedures. Significant differences between two or more ethnic groups on any of these indices in isolation or in patterns of correlations among all the units that make up the total score would suggest that the test behaves differently internally across those groups (Jensen, 1980, p. 429). Such evidence of construct bias would suggest that the scores may have a different psychological meaning across groups (Jensen, 1980, p. 533).

His analysis of data then available led Jensen (1980, pp. 585–587) to conclude that (a) White, African American, and Mexican American samples show similar internal consistency reliabilities and raw test score/chronological age correlations on mental tests; (b) social class and racial group differences in item difficulty are not consistently related to subjective judgements of the degree of "culture-loadedness" of mental test items; (c) White–African American differences are typically slightly

larger on nonverbal than on verbal tests; (d) factor analyses of mental test batteries show the same factor structure in White and African American samples; (e) the magnitude of African American–White differences (in standard deviation units) is directly related to the size of a test's *g* loadings; (f) in tests with heterogeneous item types (e.g., Stanford-Binet and Wechsler scales), the rank order of item difficulties across ethnic groups correlates above .95; (g) items that discriminate most between African American and White samples are the same ones that discriminate most between older and younger individuals within each ethnic group; and (h) insufficient evidence was available to permit firm conclusions regarding cultural bias in construct validity with less frequently researched groups, such as non-Mexican Hispanics, Native Americans, and Asian Americans.

Situational Bias

Jensen (1980) coined the term "situational bias" to describe "influences in the test situation, but independent of the test itself, that may bias test scores" (p. 377). Such factors include characteristics of the examiner (e.g., age, race, or gender), the emotional atmosphere of the testing situation, the cooperativeness and/or motivation of the examinee, and characteristics of the test instructions. As Jensen indicated, situational characteristics are not an attribute of tests themselves and thus are outside the domain of test bias itself. However, were they to affect test scores systematically as a function of ethnicity of examinees, then logically they should significantly manifest themselves in evaluations of both predictive and construct bias.

Jensen (1980) concluded that no situational aspects of the testing session contribute significantly to the test score differences among social class and ethnic groups.

In the remainder of this article, we discuss the effect of BIMT on writing, theory, and research since 1980 emphasizing research in school psychology, which has itself traditionally emphasized cognitive assessment.

FORMAL REVIEWS OF BIMT

Jensen has estimated that at the time Modgil and Modgil (1987) compiled their book, BIMT had received more than 100 published critiques, reviews, and commentaries. A computerized title search in the SocioFile database back to 1980 using the keywords "Bias in Mental Testing" yielded 52 reviews of BIMT in journals. Twenty-eight of these reviews appeared in *Behavioral and Brain Sciences* (Clarke, 1980). Two of the 52 citations are to responses by Jensen to reviews of BIMT. In 1984, Reynolds and Brown (1984b) published *Perspectives on Bias in Mental Testing,* a compilation of nine invited chapters presenting a variety of views from scholars on BIMT specifically and test bias issues generally. *Arthur Jensen: Consensus and Controversy* appeared in Modgil and Modgil's series profiling "Master-Minds" in psychology in 1987. A sizeable portion of the book is devoted to a

critique and discussion of psychological and educational implications of Jensen's writings on test bias.

Owing to BIMT's wide-ranging coverage, the emotionality over its subject matter, and the social importance/implications of its conclusions, reviews of BIMT were varied. Although many reviewers limited their comments to BIMT's actual content, others addressed a variety of tangential issues. The focus of reviews ranged from specific technical disagreements over the use of or rationale for certain mathematical formulas (e.g., see Horn & Goldsmith, 1981) to commentaries on the broader implications. The tone of reviews ranged from *ad hominem* polemics to balanced (but constructive) criticism to effusive praise. Of importance, despite the variability in reactions, no reviewer offered empirical evidence that refuted BIMT's main conclusions.

OTHER CONCLUSIONS THAT MENTAL TESTS ARE NOT BIASED

Two years after the publication of BIMT, a panel of 19 experts commissioned by the National Academy of Sciences and the National Research Council reviewed the test bias literature and concluded that well-constructed tests did not show evidence of bias against English-speaking minority groups, particularly African Americans (Wigdor & Garner, 1982). Perhaps the ultimate position paper by experts on intelligence and intelligence testing is "Mainstream Science on Intelligence" (1994), which was reprinted in the journal *Intelligence* (Gottfredson, 1997). Fifty-two professionals, all of them well known for their work on intelligence, signed the statement, reflecting an unusual degree of consensus among a group with varied theoretical views about intelligence. They said:

> Intelligence is a very general mental capability that...can be measured, and intelligence tests measure it well. They are among the most accurate (in technical terms, reliable and valid) of all psychological tests and assessments....While there are different types of intelligence tests, they all measure the same intelligence....Intelligence tests are not culturally biased against American blacks or other native-born, English-speaking peoples in the U.S. Rather, IQ scores predict equally accurately for all such Americans, regardless of race and social class. ("Mainstream Science on Intelligence," 1994, p. A17)

Owing largely to controversy engendered by Herrnstein and Murray's *The Bell Curve* (1994), the American Psychological Association (APA) assembled a Task Force of 11 highly respected psychologists with expertise in intelligence to evaluate research on intelligence and intelligence testing. Only three members were signatories to the "Mainstream Science on Intelligence" (1994) statement. About the characteristics of tests, their authoritative report (Neisser et al., 1996, p. 94) states, "none...contributes substantially to the Black/White differential in intelligence test scores," contrary to the CTBH. The consensus, then, among those actually conducting research on intelligence is that the CTBH is unsubstantiated (see also Reynolds, 1998, 1999).

PARTIAL SUPPORT FOR THE CONCLUSION THAT TESTS ARE NOT BIASED

Snyderman and Rothman (1988) analyzed 661 surveys from social scientists and educators containing questions related to the "IQ controversy." Four survey questions related to various claims about test bias, all of which had been refuted in BIMT. Overall, respondents indicated that most commonly used intelligence tests are somewhat biased against African Americans and low SES groups, particularly regarding conditions that were external to the test (BIMT's situational bias). However, the subset of respondents who reported conducting research on, or writing articles about, test bias rated the four claims as less credible (e.g., having less of a biasing effect) than did the remainder of the sample (p. 121).

PROFESSIONALS' CONTINUED CLAIMS THAT TESTS ARE BIASED

In this section, we will concentrate on claims since publication of BIMT that commonly used standardized tests are biased against ethnic subgroups. Although such claims use a variety of terminologies, they largely recycle one or more of the seven categories of the CTBH summarized at the beginning of this article. Indeed, this section is organized in terms of those categories. In a subsequent section, we will address the ways in which these criticisms violate basic principles of scientific inquiry.

Inappropriate Content

Dent (1996) provides perhaps the clearest recent statement of this aspect of the CTBH:

> It is not unrealistic to assume that [cognitive test item writers] represent middle-class America, and that the items they contribute reflect the middle-class experience. Asking an African American child who has lived in the inner-city, a Hispanic youngster, brought up in a barrio, or a refugee child who recently arrived from another country questions that reflect White American middle-class values and experiences will reveal very little about that child's cognitive ability or intellectual functioning. (p. 110)

In another instance, BIMT presents evidence that the rank order of item difficulty values (i.e., *p* values) on the Peabody Picture Vocabulary Test (PPVT) and Raven Progressive Matrices (RPM) are most similar for samples of White children who are 2 years younger than samples of African American children, which Jensen hypothesized owed to a cognitive lag in the African American group. Helms (1992), in contrast, offers this explanation:

> Maybe white children learn their own culture two or more years sooner than Black children vicariously learn White culture, but then, presumably, White

> children have more direct exposure to their own culture (and, parenthetically, the culture of [cognitive ability tests]). Because Black test takers as a group are unlikely to have the same depth of exposure to White culture as Whites have, then it is plausible that a cultural lag of some degree might be present throughout the life span. (p. 1084)

Inappropriate Standardization Samples

Dent (1996) also invokes this claim in criticizing IQ testing of African Americans, stating:

> If African Americans and other minorities are included in [standardization] samples, their scores will be dispersed throughout the distribution of obtained scores. Minority group scores will not cluster within the distribution in large enough numbers to have any influence on the norms....The largest segment of the population represented in the tryout samples and the standardization sample will be the White, middle class. This is the group which will control the greatest influence on item selection and on the norms or the standardization of a norm-referenced test. (p. 111)

Figueroa (1991) argues that since test makers face technical and financial difficulties in norming ability tests on large samples of bilingual populations, then tests should not be used in educational decision-making with such populations (Figueroa, 1991).

Examiners' and Language Bias

Hilliard (1984) evokes the "language bias" argument in criticizing BIMT for not taking into account the work of cultural anthropologists and linguists. If this were done, according to Hilliard, psychometricians would presumably be compelled to admit that "a vocabulary test or vocabulary similarities that are dependent on a common vocabulary meaning should be expected to vary by cultural group" (p. 147). Therefore, test developers cannot assume that items are apprehended or perceived in the same way by all examinees. Armour-Thomas (1992), in a selective review of articles that adopt a sociolinguistic perspective in criticizing test interpretation with different cultural groups, writes, "Testing procedures are standardized for purposes of reliability and cannot adjust for differential response biases in children that may be a function of their sociolinguistic experiences. But such constraints of the testing environment may preclude an accurate estimation of cognitive competence from children from culturally different backgrounds" (p. 557).

Inequitable Social Consequences

In responding to BIMT's conclusion that standardized tests of mental ability do not cause, but merely reflect, subgroup differences, Scarr (1981) states:

> By addressing so narrowly the scientific issues of test bias, Jensen has avoided an explicit statement of the conclusions that scream out from his pages: that blacks as a group will not succeed educationally and occupationally in this society, and their lack of success can be justified by their poor performance on tests that predict educational and occupational success...[B]y raising the specter of racial genetic differences in intelligence and by defining test bias in a narrow, psychometric way, Jensen's book conjures up images of blacks doomed to failure by their own inadequacies. (p. 330, 338)

Gould (1995, 1996) explicitly agrees that tests are not biased statistically and do not show differential predictive validity. But he claims that defining cultural bias statistically is confusing because the public is concerned not with bias in a narrow statistical sense, but whether the African American–White IQ difference occurs "because society treats blacks unfairly—that is, whether lower black scores record biases in this social sense. And this crucial question (to which we do not know the answer) cannot be addressed by a demonstration that [statistical] bias doesn't exist..." (Gould, 1995, p. 18).

Differential Predictive Validity

Some proponents of the CTBH continue to argue that tests are not valid predictors for ethnic subgroups while at the same time admitting that they do have predictive validity (e.g., Helms, 1992; Hilliard, 1984), although the specific mechanism through which this supposedly occurs differs from critic to critic. Helms (1992) argues:

> Eurocentricism likely is often intrinsic to the criteria or evaluations thereof as well as to the tests themselves. When Eurocentricism mutually characterizes tests and criteria, then significant test-criteria correlations should occur. However, the meaning of the correlations may not be that intelligence (as measured by [cognitive ability tests]) is predictive of performance (as variously measured). Instead they might merely indicate that a Eurocentric cognitive style is correlated with itself wherever one measures it. (p. 1096)

Hilliard (1984) offers a different perspective in criticizing predictive validity methodology:

> Existing psychometric predictive validity models fail to describe, control for, or account for intervening variables between IQ testing and measures of achievement. In particular, variations in the quality of instruction and variations in relevant life experiences [of minority cultural groups] are totally ignored...I have discovered no evidence to indicate that the inadequacy in the models for validity study is being examined. (pp. 165–166)

Dent (1996) cites data from Mercer (1979) showing that the correlation between the Wechsler Intelligence Scale for Children and grade point averages for K–6 students was .46 for White students but .20 for African American students. These data are then compared with statistical criteria for establishing predictive validity bias, much of which was echoed in BIMT. On the basis of this one study, Dent concludes "These data clearly indicate that...the Wechler Intelligence Scale for Children does not meet the criteria established for test fairness" (p. 113).

Qualitatively Distinct Minority and Majority Aptitude and Personality

CTBH adherents continue to claim that cognitive processes are subgroup specific and cannot be used validly even on native-born and English-speaking members of ethnic minority groups. For example, Helms (1992) argues that the similarity in test factor structures and correlation matrices across groups occurs because the measurements reflect little more than "proficiency in White culture" (p. 1084). In a similar vein, Richardson (1993) suggests that psychometricians have not settled the issue of whether or not cognitive ability tests measure general intelligence or a "Eurocentric cognitive style" (p. 565).

The best known presentation of the cognitive-difference version of the CTBH is given by Helms (1992), who claims that "European-centered" values and beliefs are characterized by "rugged individualism," "action orientation," "status and power," and "competition" (among other things). In contrast, "African-centered" values and beliefs are characterized by "spirituality," "harmony," "movement," "orality," and "social time" (among other things). Helms proposes that these different styles influence responses on cognitive ability tests.

PRESENTATIONS OF BIAS TO NONPROFESSIONAL AUDIENCES

Popular Media Presentations

An important element of the controversy over BIMT and more recent books, particularly *The Bell Curve* (Herrnstein & Murray, 1994), is vigorous media attention (e.g., Jacoby & Glauberman, 1995). Media presentations cannot be ignored because they influence professionals and lay people alike. Unfortunately, the media may distort issues, provide only partial explanations, and misrepresent or ignore professional views. One reason for the "Mainstream Science on Intelligence" (1994) statement was to counter flawed presentations: "[S]ome conclusions dismissed in the media as discredited are actually firmly supported" (p. A18). One of those media-dismissed conclusions was that tests are unbiased.

Making a case for powerful media influence in his "IQ Testing and the Media," Herrnstein (1982) stated, "A story neglected sometimes damages the truth more than a story mistold..." (p. 70). Media frequently exert influence by initially choos-

ing submissions that are critical of IQ/mental testing, making these issues controversial, and then ignoring contradictory evidence (the empirical base) or rebuttal (by experts in testing). Herrnstein (1982) chronicled his numerous dealings with media in which they both distorted his and others' views and then refused to allow them to respond to those attacks.

Media coverage of test bias, intelligence tests, and even intelligence in general is often inaccurate and itself biased. A "World News Tonight" segment on November 22nd, 1994 reflects the extent to which media slant presentation of views (see Gordon, 1997, for a detailed analysis). In the space of a few minutes, the segment attacked standard IQ tests as narrow relative to "current knowledge" (i.e., Gardner's theory of multiple intelligences; see Jensen, 1998, pp. 128–132 for a critique), as well as researchers whose work has been supported by the Pioneer Fund (an agency which funds research on racial differences). The ABC reporter then tied the Pioneer Fund to the American Eugenics Society and to Nazi death camps. The wholly unwarranted implication was, of course, that research funded by the Pioneer Fund must be tainted with eugenicist views.

Textbook Presentation of Test Bias

In an analysis of a small, but we trust not biased, sample of recent popular introductory and developmental psychology textbooks, Brown and Barbour (1999) found that of the 27 that discuss bias, 67% state that tests are biased, 11% were uncertain, and only 22% state that tests are not biased. Kalat (1999), Lefton (1997), and Weiten (1997) have particularly accurate presentations. After describing the predictive validity of tests and lower school performance of African Americans, Kalat (1999, p. 234) states, "In short, the tests show no evidence of ethnic-group bias, so continued claims of test bias could be described as blaming the messenger (the IQ tests) for the bad news." Of those claiming bias, most do not even mention differential predictive validity as a criterion for bias, but only present individual items as putatively biased. Twenty-two percent present the item, "What is the thing to do if a boy/girl much smaller than yourself starts to fight with you?" as biased although evidence cited above indicates that it is not. Eleven percent present items from the Black Intelligence Test of Cultural Homogeneity (BITCH), with no mention of its lack of validity. Many of the texts cite Helms (1992) in support of the conclusion that tests are biased, even though she offers no empirical support for her claims. Given the position of researchers that tests are not biased, Plotnik's (1996) position is particularly puzzling: "Researchers admit that IQ tests are culturally biased because they assess accumulated knowledge, which depends on environmental opportunities available in a particular culture or group" (p. 267). Plotnik cites Humphreys (1992). But Humphreys (1992) explicitly denied that tests are biased, leaving one to wonder on what Plotnik's statement actually was based.

BIAS RESEARCH IN SCHOOL PSYCHOLOGY JOURNALS SINCE BIMT

Since the publication of BIMT, studies evaluating various tests for bias have appeared frequently in the psychology literature. Since this paper focuses on test bias as it relates to school psychology, we have limited our review here to the school psychology literature.

Predictive Validity

At the time BIMT was published, the Wechsler Intelligence Scale for Children—Revised (WISC-R) was the latest edition in use. Using WISC-R Verbal IQs to predict Wide Range Achievement Test Reading subtest scores of middle versus low SES emotionally disturbed, learning-disabled, and educable mentally retarded 7- to 22-year-olds, Hale, Raymond, and Gajar (1982) found no evidence of differential predictive validity. Using WISC-R Full Scale IQs of large samples of middle and lower SES White, African American, and Mexican American first through eighth graders, Oakland (1983) found little evidence of biased prediction of California Achievement Test Reading and Math scores; validity coefficients among the three ethnic groups and two social classes were similar. Poteat, Wuensch, and Gregg (1988) found no evidence of predictive bias when they used WISC-R scores to predict California Achievement Test scores and GPAs for samples of African American and White elementary/middle school students referred for special education evaluations. Weiss and Prifitera (1995) found no evidence of bias in the prediction of Wechsler Individual Achievement Test (WIAT) scores from WISC-III Full Scale IQ scores in a large sample of White, African American, and Hispanic children.

Studies of predictive validity conducted with other cognitive tests have generally also failed to find evidence of differential predictive validity. Owing to the large number of studies to summarize, we will in each case state first the predictor and criterion measures and then the basic results. (a) Boehm Test of Basic Concepts (BTBC) Forms A and B as predictors of early school achievement on the SRA Achievement Series and first grade basal-reader placement in samples of White and Mexican American children. The only significant bias was that BTBC Form B overpredicted reading performance of Mexican Americans (Reynolds & Piersel, 1983); (b) Two McCarthy Scales of Children's Abilities subtests, Lee-Clark Readiness Test, Mathematics and Language subtests of the Tests of Basic Experiences, Preschool Inventory-Revised Edition, and Metropolitan Readiness Tests as predictors of Metropolitan Achievement Test (MAT) scores of groups of African American and White children. Out of 112 statistical analyses, 13 produced instances of overprediction of MAT scores for the lower-scoring African American group (Reynolds, 1980, 1983); (c) Kaufman Assessment Battery for Children (K-ABC)

global cognitive scores as predictor of Comprehensive Tests of Basic Skills (CTBS) scores for English-speaking Mexican American and White samples of fifth and sixth graders of similar socioeconomic status. K-ABC scores predicted CTBS scores less well for the Mexican American sample than for the White sample (Valencia & Rankin, 1988); (d) Raven Coloured Progressive Matrices (CPM) and the Nonverbal Test of Cognitive Skills (NTCS) as predictors of GPAs and three California Achievement Test subtest scores for samples of White and Mexican American second and third graders. The CPM exhibited slope and/or intercept bias on three of the four criterion variables, whereas the NTCS demonstrated slope bias only on GPA (Emerling, 1990); (e) a kindergarten screening battery as predictor of Stanford Achievement Test (SAT) total battery scores in large samples of Native American and White students in kindergarten, first, third, and fourth grades. Use of common regression lines underpredicted Whites' scores and overpredicted Native Americans' scores (Stone & Gridley, 1991); (f) Differential Abilities Scales (DAS) ability subtest scores as predictors of DAS basic number skills achievement scores in samples of Asian American and White children. Systematic errors in prediction increased with ability level. Use of a common regression line led to overprediction of the White group's scores and underprediction of the Asian Americans' scores (Stone, 1992).

Content Validity of Individual Items

Analysis of individual items continues to be a popular method for investigating bias in the internal characteristics of mental tests. Mishra (1982) used the likelihood ratio chi-square statistic to investigate item bias in 79 items from the Information, Similarities, and Vocabulary subtests of the WISC-R for White and Native American fourth and fifth graders. Fifteen of the 79 items were biased against the Navajo participants. Comparing the rank order of difficulty on items on the Boehm Test of Basic Concepts standardization data for children of different socioeconomic levels, Silverstein, Belger, and Morita (1982) found no evidence of internal bias. Children mastered basic concepts in the same temporal order while differing only in the rate of mastery. In an evaluation of item bias in the McCarthy Scales of Children's Abilities in samples of White and Mexican American 5–8-year-old children, Murray and Mishra (1983) concluded that only three verbal items were biased.

Evaluation of item difficulty patterns in the Verbal subtests of the WISC-R for White, African American, Hispanic, and Bermudian 7–10-year-old children identified a small number of items as differentially difficult for particular groups, but the item difficulty curves for all groups were remarkably parallel (Sandoval, Zimmerman, & Woo-Sam, 1983). Koh, Abbatiello, and Mcloughlin (1984) analyzed responses from 360 test protocols of Chicago school children to examine content bias in seven WISC items singled out by Judge J. F. Grady in his opinion in the PASE (Parents in Action in Special Education) case as being culturally biased

against African American children. No significant differences were found in the percentage of students passing the items for Whites and African Americans, and error analyses showed no significant "cultural" differences between White and African American participants. Pugh and Boer (1989) examined 10 questions with culturally specific content from the Wechsler Adult Intelligence Scale—Revised (WAIS-R) Information subtest in a sample of 95 17–64-year-old Canadian participants. Rank order and chi-square analyses showed that accuracy scores (p values) for these questions differed significantly from those of the WAIS-R normative sample for the majority of the items.

Analyzing data from the entire standardization sample of the Peabody Picture Vocabulary Test—Revised (PPVT-R) using race, item, and total score intercorrelations for each item on Forms L and M, Reynolds, Willson, and Chatman (1984) found few instances of bias. Argulewicz and Abel (1984), examining item bias on PPVT-R Forms L and M for samples of White and Mexican American children in Grades 1–4, found evidence of bias in occasional items. However, the lack of a discernible pattern in biased items coupled with high reliability indices in both groups suggests that PPVT-R content bias is minimal. Using a linguistic analysis to analyze error responses, Anderson and Morris (1989) found no evidence of internal bias on the Woodcock Language Proficiency Battery in a sample of 12–14-year-old rural learning-disabled and normally achieving students. Willson, Nolan, Reynolds, and Kamphaus (1989), using a partial correlation technique to detect differential item functioning in items from the K-ABC, found that 23 items were biased against African Americans and 13 items were biased against Whites. The authors concluded, however, that elimination of these items would have little effect on race differences in mean total scores. Comparisons of item-group partial correlations on the K-ABC Mental Processing and Achievement scales for English-speaking Mexican American and White fifth and sixth graders revealed that 17 out of 120 items of the Mental Processing Scales showed evidence of bias, mostly against Mexican Americans, whereas 58 out of 92 Achievement scale items showed bias against Mexican Americans (Valencia, Rankin, & Livingston, 1996).

Construct Validity

The Wide Range Achievement Test (WRAT) met minimum requirements of internal consistency reliability for a large sample of Mexican American fourth and fifth graders (Mishra, 1981a). Internal consistency reliability estimates for the Raven Coloured Progressive Matrices (CPM) were high for both White and Mexican American third grade males from low SES backgrounds, indicating no internal consistency bias (Valencia, 1984).

Factor analytic methods are also frequently used to assess the equivalence of constructs measured in two or more groups on the same test. Factor analysis of scores for Whites and Mexican American 5–7-year-old children on 18 tests of the

McCarthy Scales of Children's Abilities yielded similar factors with both groups and no evidence of construct bias (Mishra, 1981b). Tests for factorial equivalence of scores on 10 subscales of the WISC-R for large samples of middle and low SES children indicated that the subscales measured equivalent constructs in the two groups (Hale, 1983). Reynolds and Piersel (1983) examined the cross-group factorial congruence of the Boehm Test of Basic Concepts (BTBC) Forms A and B across White and Mexican American children and found no significant construct bias. Using a multi-sample confirmatory factor analysis approach to detect construct bias in the K-ABC for large samples of White and African American 7–12-year-olds, Keith, Fugate, DeGraff, Diamond, Shadrach, and Stevens (1995) found no differences in 7–8-year-old children and small differences, apparently caused by measurement error, in 9–12-year-old children.

Situational and Miscellaneous Sources of Bias

Mishra (1980) analyzed data from the administration of two WISC verbal subtests and the Raven Progressive Matrices by two White and two Mexican American examiners on American and Mexican American third graders. Participants' performance was unaffected by examiners' ethnicity on two of the three tests, but Mexican American children scored significantly higher when the third test was administered by Mexican American examiners. Mishra (1983) gave 36 Stanford-Binet protocols of 5–8-year-old participants to four groups of examiners for scoring, but varied the amount of information about the participants' ethnicity and IQ. No bias in the process of scoring was found. We remind the reader that situational bias, if found, is independent of the test itself.

Some studies would have had potential implications for indirectly testing the "standardization fallacy" (Jensen, 1980), if they did not have serious methodological problems. For example, Oplesch and Genshaft (1981) compared 20 bilingual Puerto Rican first through third graders' scores on the WISC-R with scores on a Spanish language version of the WISC (Escala de Inteligencia Wechsler Para Ninos). The researchers found no significant differences between the Full Scale IQs across the two tests. Unfortunately, conclusions from this study are limited by a small sample size and a confounding of test translation with form of the test used (WISC versus WISC-R). Sharpley and Stone (1985) administered the PPVT-R Forms L and M (normed on Americans) to 410 nonreferred Australian children. Although the Australian sample generally scored lower, no significant differences occurred in mean raw scores in age cohorts across the American standardization sample and the Australian sample. Unfortunately, the authors did not conduct any predictive or item bias analyses across the two samples.

Too few studies of situational and similar sources of bias have appeared in school psychology journals to permit firm conclusions. For a more extended discussion of issues related to cultural minorities and the testing session, see Frisby (1999a, this issue).

Summary of Research Since BIMT

The large number of predictive validity studies from school psychology journals shows little evidence of any bias and no evidence of flagrant bias against American, English-speaking minority groups. When predictive bias is found, it typically involves intercept differences, which are to be expected with tests of less than perfect reliability. In many such cases, use of a common regression line results in overprediction of the lower scoring group and underprediction of the higher scoring group, confirming BIMT's conclusion.

Construct validity studies at the item level or of the entire test through factor analysis yield the following conclusions: (a) bias is occasionally found in certain items; (b) items often judged by "armchair" observation (face validity) to be biased prove to be unbiased when subjected to empirical analysis (e.g., Jensen & McGurk, 1987; Sandoval & Miille, 1980); (c) in most tests, more items show no evidence of bias than show evidence of bias; (d) observed item bias is often unsystematic, such that patterns in item content or specific groups of examinees penalized cannot be discerned; (e) observed item bias is too small to explain the size of group differences in mean total scores; and (f) factor analytic methods reveal factor equivalence across groups and thus no construct bias. For more detailed discussions of statistical problems in item bias research and other bias studies, see Camilli and Shepard (1987, 1994) and Reynolds, Lowe, and Saenz (1999).

EMPIRICAL AND CONCEPTUAL DEVELOPMENTS IN TEST BIAS SINCE BIMT

In 1980, the year of publication of BIMT, over 180 researchers, test developers, and test users were invited to attend the Third Annual Johns Hopkins University National Symposium on Educational Research, held in November. The purpose of the symposium was to synthesize research, propose new guidelines for practice, and identify new research problems related to the topic of test bias. Contributions from 17 distinguished presenters at the symposium were published in the *Handbook of Methods for Detecting Test Bias* (Berk, 1982), which describes methods employed by contemporary test publishers to remove bias from achievement, aptitude, and intelligence tests used in public schools, college admissions, and professional certification examination contexts.

Item Bias Detection

According to Jensen (1987), space limitations prevented inclusion in BIMT of a more detailed discussion of Item Response Theory (IRT)-based methods for detecting item bias. IRT methods overcome several shortcomings of item analysis methods based on classical test theory (e.g., Hambleton & Swaminathan, 1985). Based on IRT, the analysis of item characteristic curves (ICC) has become the most frequently used method for detecting item bias in very large samples. An ICC is a

monotonically increasing function that describes the relationship between an increase in the probability of a correct response on an item and an increase in the trait that underlies item performance (Hambleton, Swaminathan, & Rogers, 1991). According to Osterlind (1989), ICCs for two or more groups on the same item are equated, then plotted together as a function of item difficulty (probability of success) and examinee ability. Good ICCs resemble the "S" shaped ogive of the normal distribution, whereas poor items depart from this shape. An item is unbiased if the probability of success on the item is the same for equally able examinees regardless of their group membership (Osterlind, 1989). When the same ICCs for two or more groups are superimposed and do not overlap completely, the area between the equated ICCs indicates the degree of bias present in the item. According to Osterlind (1989), IRT procedures involve complex statistics, require sophisticated computer programs, and work best with very large sample sizes from each subpopulation in order to produce stable item parameter estimates. Unfortunately, obtaining large samples is difficult with some subgroups. For these reasons, this family of methods is used mainly by test development companies that have access to sufficient computer resources and sample sizes. The result is that major contemporary mental tests show no bias against American-born, English-speaking subgroups. A recent presentation of modern methods for detecting item bias can be found in Camilli and Shepard (1994).

Consequential Validity

Cole and Moss (1989), reflecting a position that expands the concept of test bias beyond its purely statistical/technical meaning as detailed in BIMT, discuss "extra-validity" issues, defined as "the purposes for which a test is used, the extent to which those purposes are accomplished by the actions taken, the various side effects or unintended consequences, and possible alternatives to the test that might serve the same purpose" (p. 213).

In particular, Messick (1989, 1996) argues that the traditional conception of content, criterion, and construct validity is incomplete and should be expanded to incorporate extra-validity concerns such as *consequential validity*. Consequential validity would include politics, values, and culture in considering the full context of test interpretation and test use (see Reckase, 1998, for an example).

However, issues of consequential validity largely relate to the fairness of the application of tests. Including them as an aspect of overall test validity would be retrogressive, reintegrating considerations of fairness with those of test bias. An important contribution of BIMT, which test critics ignore, is its distinction between bias and fairness. Writers continue to add new meanings to the concept of test bias. Although the APA Task Force Report on Intelligence (Neisser et al., 1996) states that cognitive tests show no predictive bias in the statistical sense, they nevertheless charge tests with "outcome bias" because of differential consequences for different groups. Reschly (1997) explicitly conjoins bias and unfairness: "Assessment that

does not result in effective interventions...may be biased or unfair...if children and youth with minority characteristics are differentially exposed to ineffective programs as a result of assessment activities" (p. 438).

Although consequential validity continues to be studied (Reynolds, in press), Lees-Haley (1996) severely criticizes it as degrading the more objective features of measurement that form the foundation of science.

Cross-Cultural Test Bias

Although BIMT included a chapter on "culture-reduced" tests, the chapter did not discuss in detail bias issues related to American test translations in different countries, the testing of non-English or limited-English speakers, or test interpretation in cross cultural (cross-country) comparisons. Van de Vijver and Poortinga (1994, 1997) argue that psychometric analyses of item bias should be supplemented with cross-cultural analyses of bias in two additional areas: "construct bias" and "method bias." According to these authors, construct bias occurs when "test authors from various societies use definitions of the concept under study that do not fully overlap" (van de Vijver & Poortinga, 1997, p. 30). Method bias occurs "when a cultural factor that is not relevant to the construct studied affects most or all items of a test in a differential way across the cultures studied" (p. 30). Such biases can occur when groups from different countries are tested under different testing conditions or are differentially familiar with response procedures. Since these concerns are outside bias within a given overall culture, they are not discussed here in detail. More detailed treatments are in Bracken and Barona (1991), Frisby (1999b, this issue), Geisinger (1994), Hamayan and Damico (1991), Hambleton (1993), Hambleton and Bollwark (1991), Irvine and Berry (1988), and Lam (1993).

WHERE DO WE GO FROM HERE? TEST BIAS AND PRINCIPLES OF SCIENTIFIC INQUIRY

Given the overall current state of research and theory on bias and methodology for test construction and evaluation, we argue that contemporary defenders of the CTBH as an explanation for mean score differences on mental tests between American-born, English-speaking subgroups have not adhered to rudimentary principles of science. Below we present these principles, and the manner in which they have been violated.

1. *Scientific terms should be operationally defined in order to promote clarity in discussion, debate, and problem solving.* Consensus in conceptual definition is not needed in order to study a phenomenon. After all, researchers continue to study intelligence despite a lack of agreement over common definition of what it is (e.g., Sternberg & Detterman, 1986). But operational definitions of the manipulation or measurement of a concept are necessary. Operationally, a biased test is one that yields differential predictive validity. Many defenders of the CTBH appear to use

an implicit definition: A biased test is one that yields subgroup differences. The first definition is in keeping with the everyday meaning of bias: "a particular tendency or inclination, esp. one which prevents unprejudiced consideration of a question" (Barnhart, 1951), whereas the second prejudges tests as biased if they show group differences.

Rosenbach and Mowder (1981) observe that any clinician with a hunch that poorly performing test takers can do better on tests can accuse tests of bias from a "clinical" or "intuitive" perspective. Entire texts have been written with this implicit definition of bias in mind (e.g., Armour-Thomas & Gopaul-McNicol, 1998). Helms (1992) concedes that cognitive tests show no bias from a purely statistical perspective, yet accuses tests of bias from a "culturalist" perspective, presumably because they have not demonstrated "cultural," "functional," or "linguistic" equivalence across groups, concepts which she fails to operationalize.

BIMT should have convinced professionals that a group difference in test scores is not an acceptable operational definition of test bias. The reason is clear, since a group difference definition could conceivably allow someone to accuse reading comprehension tests of being biased against illiterates (Cameron, 1988). Unfortunately, Jensen's attempt has not been particularly successful, as seen in several places in this article.

BIMT should also have convinced scientists and other professionals that bias cannot be identified from "armchair" face validity analysis of items (e.g., see Jensen & McGurk, 1987; Sandoval & Miille, 1980). However, professional educators sometimes reveal their own confusion. Consider a recent televised interview (October 14, 1997) in Austin, Texas: A reporter asked a ranking Texas Education Agency (TEA) testing official to comment on a recent lawsuit alleging that the TEA competency test, which must be passed to receive a high school diploma, is biased against minorities. The allegations themselves were based on mean subgroup differences. The official replied, essentially, that the test could not be biased because members of minority groups read test items prior to their use and identified biased ones, which were then deleted.

Understandably, charges that a test is biased are interpreted generally as meaning that the test is bad. Therefore, any test critic can use his or her own definition of bias to try to convince the lay public that something is wrong with a test.

Lack of operational definition of scientific terms has the unfortunate consequence of leading test defenders and critics to argue *past* each other instead of *with* each other. Furthermore, violation of this scientific principle may lead to violations of other scientific principles.

2. *Theories should be consistent with the data they are intended to explain.* Many defenders of the CTBH propose hypotheses and theories that simply are not congruent with extant empirical data. Thus, Helms (1992) writes as if a homogeneous "White" or "Eurocentric" culture explains group psychological differences, underlies the orientation of test developers, and saturates item content on mental tests. But evidence indicates that Whites are not homogeneous culturally or ethnically,

and that mean IQ differences can be found among "White" subgroups (see Eysenck, 1984). Armour-Thomas's (1992) claim that some "cultural" groups have greater social opportunities or experiential advantages over other groups presumably explains group differences in test scores and implies that standardized cognitive tests are invalid when used with diverse populations. But the mean mental test scores of low SES Whites exceeds that of high SES African Americans (Coleman et al., 1966; Scarr-Salapatek, 1971; Shuey, 1966) and African American–White IQ differences stabilize at one standard deviation even at the highest SES levels (Herrnstein & Murray, 1994).

Perhaps most importantly, reliable evidence of equivalent predictive validity and higher mental test scores by Asian Americans (e.g., Reynolds et al., 1999) directly contradicts the CTBH. Francis Bacon's "crucial experiment" (e.g., Brown & Reynolds, 1984), which tests contradictory predictions from two theories, is a strong method for deciding between two theories. The psychometric position, that mental tests measure "g" equally for all subgroups, and the CTBH make opposing predictions. Much data, including evidence that subgroups differ most on abstract tests such as Raven's progressive matrices and digits backward (e.g., Gordon, 1984; Jensen, 1980, 1984b), clearly support "g" theory and contradict the CTBH.

3. *A scientific theory should yield clear testable predictions.* One contribution of BIMT was its specification of empirical tests of the CTBH. If mental tests are biased against a particular group, then the following predictions should be confirmed: (a) items would show a different rank order of item difficulties for different groups; (b) item-total score correlations would differ significantly among groups; (c) item × group interaction terms in ANOVA designs would achieve statistical significance; (d) ICCs for different groups would show different curves; and generally (e) tests should show differential predictive validity for different groups. Given that these results do not occur, the CTBH has effectively been disconfirmed.

In contrast, the hallmark of the CTBH (Helms, 1992; Hilliard, 1984) is attribution of group differences to a vague "cultural" variable that, however real to test critics, is so elusive that it has managed to escape detection by any method that can be devised by modern psychometrics. Helms's (1992) suggestion that the psychometric methods are themselves biased renders the hypothesis wholly untestable because she can account for any research result, whatever it is. The CTBH is characterized by an absence of specific predictions related to test data, statistical methods that would test these predictions, or any objective criteria that would enable its scientific evaluation.

4. *Science accepts the principle of parsimony (Ockham's razor).* Science accepts the principle that, other things being equal, the simplest explanation is the best. The simplest scientific explanation of an event is that it owes to chance, and unless evidence indicates otherwise, science does not reject that null hypothesis. Given the claim that tests are biased, the null hypothesis is that they are *not* biased. That hypothesis should be rejected only on the basis of contradictory empirical evidence.

Any argument of test bias logically reduces to one that test scores should underpredict minority group members' "real" abilities. Therefore, all bias in tests must ultimately manifest itself as differential predictive validity reflected in significantly lower accuracy of prediction for minority groups relative to the majority. But overwhelming evidence supports the null hypothesis that tests do not have differential predictive validity.

5. *Critical and supportive evidence should be empirical and relevant. The essential task of CTBH defenders is to present relevant evidence that refutes the conclusion from BIMT and subsequent research that mental tests have equivalent predictive validity for different groups.* In order to accomplish this, critics must demonstrate that tests show systematic error when administered to different groups, and that such error operates to the *detriment* of minority groups.

Whether or not test defenders fail to offer "remedies" for racial differences (Scarr, 1981), fail to respect research by cultural linguists and anthropologists (Hilliard, 1984), fail to address societal concern over racial discrimination in general in dealing with test bias (Gould, 1995, 1996), use massive empirical data to cover historic prejudice and racism (Richardson, 1995), fail to recognize that subgroups have different cognitive/linguistic processes (Figueroa, 1991; Helms, 1992), do not address special education programs that are inadequate (Reschly, 1997), or fail to include more African Americans in test standardization samples (Dent, 1996) may be interesting issues, but ultimately are irrelevant to scientific evaluation of test bias. Much of the emotionalism in debates on test bias can be attributed to such irrelevant arguments by CTBH defenders.

6. *A scientific theory should be congruent with other theories and data.* The CTBH is ultimately rooted in the "specificity doctrine" (Jensen, 1984b): mental abilities comprise "a repertoire of specific items of [learned] skills and knowledge" and are measured by mental tests that "measure nothing other than some selected sample of the total repertoire of knowledge and skills deemed important by the test constructor" (p. 94).

A rival hypothesis to the specificity doctrine is "g" theory, first articulated by Spearman (1927), and researched extensively by contemporary psychologists. It accounts for the relative size of White–African American mean differences across groups of diverse tests in terms of each test's *g* loading, with larger mean differences occurring with tests with larger *g* loading. Much empirical data supports this explanation.

Even if the CTBH acceptably accounted for data on group differences in mental test performance, which it does not, it would fail on the grounds of lack of fit with other data on intelligence. That is, it exists in isolation from other data, whereas *g* theory accounts for a massive amount of the literature on measured intelligence (see Jensen, 1998).

7. *Neither theories nor evidence should be evaluated on the basis of popularity or consistency with particular sociopolitical positions.* Political and social attitudes range from the "far left" to the "far right" (Kerlinger, 1984b). A legitimate question

to ask is whether scientific research influences, or is influenced by, shifting political/popular social ideologies in which it is imbedded (Jensen, 1984c). For example, Reynolds and Brown (1984a) review particularly odious manifestations of "scientific racism" in early writing on group differences and human measurement. The point of their review is to help readers understand reasons for objections to testing that persist in modern times. In particular, many African American professionals exhibit hostility toward testing (Gordon, 1987). Regardless of emotional reactions, conclusions that tests are unbiased should be evaluated with respect to whether or not they withstand empirical scrutiny, and not according to the popularity of their message.

IMPLICATIONS FOR SCHOOL PSYCHOLOGY

As school psychologists are frequent users of standardized mental tests, they should rightly be concerned about the tests' psychometric properties, including the obviously important issue of bias. A major message of this paper is that the overwhelming body of research confirms the equivalent validity of those tests with culturally diverse groups. Thus, school psychologists should be confident in the tests' value. Of course, the tests have limitations; their predictive validity is hardly perfect, and they doubtless do not adequately measure the abilities of some individual examinees, for example.

Given tests' limitations, school psychologists may also rightly consider the use of other assessment procedures. Alternative methods have some advantages over "traditional" cognitive tests (Shinn, 1998). As such, they can be useful supplements to accepted standard practice. Unfortunately, some professionals propose that school psychologists abandon standardized instruments with good psychometric properties in favor of assessment procedures that either are based on intelligence theories with no long history of research (e.g., Armour-Thomas, 1992) or have little or unknown psychometric bases (e.g., Armour-Thomas & Gopaul-McNicol, 1998; Mercer & Lewis, 1978). Suzuki, Meller, and Ponterotto (1996, p. 680) state that, "the continued development of alternative measures and procedures will facilitate movement of the profession toward more culturally-sensitive assessment practices." However, the impetus to move the practice of school psychology toward alternative methods for minority groups rests largely on the unfounded belief that current standardized tests are biased.

Little research is available on possible bias in the proposed alternative assessment measures, suggesting that their use is premature at best and perhaps contrary to best practice. Methods such as performance-based assessment are generally less reliable and valid than standardized tests (e.g., Rotberg, 1995). Since reliability of assessment procedures is negatively correlated with bias (i.e, the lower the reliability, the greater the likelihood of bias; see Linn & Werts, 1971), performance-based assessment instruments are likely to be biased to some extent (see Braden, 1999, this issue). Professional organizations, including the NASP (1993), have issued

cautions regarding performance-based instruments because of their limited empirical base (see Braden, 1999, this issue).

As Jensen (1980) cautioned, "Claims of test bias and of the unfair use of tests cannot be ignored by psychologists. Such claims must be objectively investigated with all of the available techniques of psychometrics and statistical analysis" (p. ix). Jensen then states explicitly that where bias is found, the test either should not be used on the subgroups for which it is biased or should be revised so as to eliminate the bias. But he goes on to say: "Before the use of tests is rejected outright, however, one must consider the alternatives to testing—whether decisions based on less objective means of evaluation (usually educational credentials, letters of recommendation, interviews, and biographical inventories) would guarantee less bias and greater fairness for minorities than would result from the use of tests" (p. ix). More recently, others have adopted the same position (e.g., Daniel, 1997).

When tempted to abandon standardized tests, school psychologists should first ensure that alternatives have equivalent psychometric properties. Furthermore, as numerous authors have indicated, those who urge that use of mental tests be abolished have the responsibility to present equally reliable and valid alternative methods of assessment. Unfortunately, as Suzuki and Valencia (1997) warned, research on test bias is decreasing at a time when it should be expanded to cover other subgroups and assessment techniques.

CONCLUSIONS

The major conclusion of this article can be stated confidently: Empirical evidence overwhelmingly supports the conclusion that well-developed, currently-used mental tests are of equivalent predictive validity for American-born, English-speaking individuals regardless of their subgroup membership. Individual items in new tests are now routinely evaluated for bias in the development stage. Test developers use empirical approaches to detect differential item functioning across nominal groupings (typically by ethnicity and by gender) of examinees. Item-response theory (e.g., Hambleton, Swaminathan, & Rogers, 1991; Osterlind, 1989) is consistently applied in test development, as in computer-adapted tests, and virtually ensures that items are not biased. The presence of biased items is quite unusual in modern tests (Reynolds et al., 1999). Increasingly, test developers are advertising their tests as unbiased for ethnicity, race, or gender. Since test companies have access to large and representative samples, any apparent item bias that is discovered in a subsequent research study typically owes to the study's small and unrepresentative sample. Although early intelligence instruments had a variety of psychometric limitations, current frequently used mental tests are perhaps the most carefully constructed, standardized, and evaluated of all psychological tests. In many ways, they serve as models to guide practice in other areas of test development.

Although she was addressing group differences in IQ rather than test bias specifically, we agree with Gottfredson's (1994) call for scientific integrity:

> All that is required is for scientists to act like scientists—to demand, clearly and consistently, respect for truth and for free inquiry in their own settings, and to resist the temptation to win easy approval by endorsing a comfortable lie. (p. 59)

To suggest that science operates separately and distinctly from society and its politics is naive. The controversies surrounding nuclear power, germ warfare, cloning, and fetal tissue implants attest to the fact that science does not take place in a vacuum and therefore should not be exempt from scrutiny by society at large. We need to recognize that the products of science (as conceived by scientists and received by the general public) are subject to moral, ethical, and personal examination. How a particular scientist interprets data is subject to many of these same factors. As an affected party, society must take part in decisions about the application of scientific knowledge in terms of desired outcomes.

However, society can appropriately influence policy and decision-making only to the extent that it is accurately informed about scientific evidence. Before science has had the opportunity to provide technical information to the public in an understandable form, however, misinformation pertaining to test bias has often been presented by numerous sources, many of which are themselves biased. Because of the powerful influence of the media and the extent of existing misunderstanding, scientists and practitioners both should consider their responsibility to provide relevant information (based on sound scientific practice) to their own profession, their students, and the general public. Professionals in school psychology, individually and collectively, need to present the methodological issues and empirical evidence on socially relevant issues such as bias in mental testing in an informed and objective manner. Having done this, practice should follow sound empirical knowledge. Little is of greater value to a professional than a strong heuristic with clear empirical support.

We agree with Frisby's concern about *indiscriminate talk*, expressed in the introduction to this series of articles. The truism that everyone is entitled to his or her opinion is matched by the truism that not all opinions are equal. In science, certainly not all opinions are equal, and peer review should weed out those opinions that are empirically unjustified or logically inconsistent. But as citations contained in this article indicate, the weeds are endangering the orderly garden. Publication in highly regarded professional journals of articles whose claims are contradicted by empirical evidence undermines the credibility of both our journals and psychology's very claim to be scientific. Can we expect others to take us seriously when we do not take each other seriously? If, as Hull (1988) has argued persuasively, science operates much as does evolution, in the long run, only the better data and theories will survive. However, waiting for the evolutionary pro-

cess to work is a painfully slow process. In the short run, a version of Gresham's Law (bad money drives out good) may operate: Bad ideas repeated often enough and loudly enough by those with "names" may, through tenacity and authority, drive out good ideas. Test validity is controversial perhaps only because of its impact on social decision making. But sociopolitical expediency is a poor justification for unsupportable ideas. Bias in belief may be more damaging ultimately than bias in mental testing.

REFERENCES

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

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

Anderson, P. L., & Morris, P. D. (1989). Use of the Woodcock Language Proficiency Battery with culturally variant learning-disabled students. *Psychology in the Schools, 26,* 130–138.

Argulewicz, E. N., & Abel, R. R. (1984). Internal evidence of bias in the PPVT-R for Anglo-American and Mexican-American children. *Journal of School Psychology, 22,* 299–303.

Armour-Thomas, E. (1992). Intellectual assessment of children from culturally diverse backgrounds. *School Psychology Review, 21,* 552–565.

Armour-Thomas, E., & Gopaul-McNicol, S. (1998). *Assessing Intelligence.* Thousand Oaks, CA: Sage.

Barnhart, C. L. (Ed.). (1951). *The American college dictionary.* New York: Harper.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore, MD: Johns Hopkins University Press.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119–132.

Braden, J. P. (1999). Performance assessment and diversity. *School Psychology Quarterly, 14,*304–326.

Brown, R. T., & Barbour, D. E. (1999, March). *Biased* Tests or *Biased* Texts? Annual convention of the Southeastern Psychological Convention, Savannah, GA.

Cameron, R. G. (1988). Issues in testing bias. *College and University, 64,* 269–279.

Camilli, G., & Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics, 12,* 87–99.

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

Clarke, A. M. (1980). Unbiased tests and biased people. *Behavioral and Brain Sciences, 3,* 337–339.

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

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Daniel, M. H. (1997). Intelligence testing: Status and trends. *American Psychologist, 52,* 1038–1045.

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

Emerling, F. (1990). An investigation of test bias in non-verbal cognitive measures for two ethnic groups. *Journal of Psychoeducational Assessment, 8,* 34–41.

Eysenck, H. J. (1984). Effect of race on abilities and test scores. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 253–291). New York: Plenum.

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

Frisby, C. L. (1999a). Culture and test session behavior: Part I. *School Psychology Quarterly, 14*, 263–280.

Frisby, C. L. (1999b). Culture and test session behavior: Part II. *School Psychology Quarterly, 14*, 281–303.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304–312.

Gordon, R. A. (1987). Jensen's contributions concerning test bias: A contextual view. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy* (pp. 77–154). New York: Falmer Press.

Gordon, E. W., & Bhattacharya, M. (1994). Race and intelligence. In R.J. Sternberg (Ed.), *Encyclopedia of human intelligence* (Vol. 2, pp. 889–899). New York: Macmillan.

Gordon, R. A. (1997). *How smart we are about what we broadcast: An open letter to ABC News*. Unpublished paper, available from Pioneer Fund, New York, NY.

Gordon, R. A. (1987). Jensen's contributions concerning test bias: A contextual history. In S. Modgil & C. Modgil (Eds.). *Arthur Jensen: Consensus and controversy* (pp. 77–154). New York: Falmer.

Gottfredson, L. S. (1994). Egalitarian fiction and collective fraud. *Society, 31*(3), 53–59.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*,13–23.

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: BasicBooks.

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

Hale, R. L. (1983). An examination for construct bias in the WISC-R across socioeconomic status. *Journal of School Psychology, 21*, 153–156.

Hale, R. L., Raymond, M. R., & Gajar, A. H. (1982). Evaluating socioeconomic status bias in the WISC-R. *Journal of School Psychology, 20*, 145–149.

Hamayan, E. V., & Damico, J. S. (Eds.). (1991). *Limiting bias in the assessment of bilingual students*. Austin, TX: Pro-Ed.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9*, 57–68.

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., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

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

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

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

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: Plenum.

Hoffman, B. (1962). *The tyranny of testing*. New York: Crowell-Collier.

Horn, J., & Goldsmith, H. (1981). Reader be cautious: A review of Bias in Mental Testing. *American Journal of Education, 89*, 305–329.

Hull, D. L. (1988). *Science as a process*. Chicago: University of Chicago Press.

Humphreys, L. G. (1992). Ability testing. *Psychological Science, 3*, 271–274.

Irvine, S. H., & Berry, J. W. (1988). *Human abilities in cultural context*. New York: Cambridge University Press.

Jacoby, R., & Glauberman, N. (Eds.). (1995). *The Bell Curve debate: History, documents, opinions*. New York: Times Books.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39*, 1–163.

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

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

Jensen, A. R. (1984b). Test validity: *g* versus the specificity doctrine. *Journal of Social and Biological Structures, 7*, 93–118.

Jensen, A. R. (1984c). Political ideologies and educational research. *Phi Delta Kappan, 65*, 460–462.

Jensen, A. R. (1987). Differential psychology: Toward consensus. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy*, (pp. 353–399). New York: Falmer Press.

Jensen, A. R. (1994). Race and IQ scores. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (Vol. 2, pp. 899–907). New York: Macmillan.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Jensen, A. R., & McGurk, F. C. J. (1987). Black-white bias in "cultural" and "noncultural" test items. *Personality and Individual Differences, 8*, 295–301.

Kalat, J. W. (1999). *Introduction to psychology* (5th ed.). Pacific Grove, CA: Brooks/Cole.

Keith, T. Z., Fugate, M. H., DeGraff, M., Diamond, C. M., Shadrach, E. A., & Stevens, M. L. (1995). Using multi-sample confirmatory factor analysis to test for construct bias: An example using the K-ABC. *Journal of Psychoeducational Assessment, 13*, 347–364.

Kerlinger, F. N. (1984). *Liberalism and conservatism: The nature and structure of social attitudes*. Hillsdale, NJ: Erlbaum.

Koh, T., Abbatiello, A., & Mcloughlin, C. S. (1984). Cultural bias in WISC subtest items: A response to Judge Grady's suggestion in relation to the PASE case. *School Psychology Review, 13*, 89–94.

Lam, T. C. M. (1993). Testability: A critical issue in testing language minority students with standardized achievement tests. *Measurement and Evaluation in Counseling and Development, 26*, 179–191.

Lawler, J. M. (1978). *IQ, heritability, and racism*. New York: International Publishers.

Lees-Haley, P. R. (1996). Alice in validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51*, 981–983.

Lefton, L. A. (1997). *Psychology* (6th ed.). Needham Heights, MA: Allyn & Bacon.

Linn, R. L., & Werts, C.E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*, 1–4.

Mainstream Science on Intelligence. (1994). *Wall Street Journal*, pp. A17–A18.

Mercer, J. R., & Lewis, J. (1978). *System of multicultural pluralistic assessment. SOMPA*. New York: Psychological Corporation.

Mercer, J. R. (1979). Expert testimony in the Larry P. trial *(Larry P. v. Riles)*. 495 F. Supp., Northern District CA.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1996). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Mishra, S. P. (1980). The influence of examiners' ethnic attributes on intelligence test scores. *Psychology in the Schools, 17*, 117–122.

Mishra, S. P. (1981a). Reliability and validity of the WRAT with Mexican-American children. *Psychology in the Schools, 18*, 154–158.

Mishra, S. P. (1981b). Factor analysis of the McCarthy Scales for groups of White and Mexican-American children. *Journal of School Psychology, 19*, 178–182.

Mishra, S. P. (1982). The WISC-R and evidence of item bias for Native-American Navajos. *Psychology in the Schools, 19*, 458–464.

Mishra, S. P. (1983). Effects of examiners' prior knowledge of subjects' ethnicity and intelligence on the scroring of responses to the Stanford-Binet scale. *Psychology in the Schools, 20*, 133–136.

Modgil, S., & Modgil, C. (Eds.). (1987). *Arthur Jensen: Consensus and controversy*. New York: Falmer Press.

Murray, A. M., & Mishra, S. P. (1983). Interactive effects of item content and ethnic group membership on performance on the McCarthy scales. *Journal of School Psychology, 21*, 263–270.

National Association of School Psychologists. (1993, April). School Psychologists' involvement in the role of assessment. Position statement adopted by NASP Delegate Assembly.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Oakland, T. (1983). Concurrent and predictive validity estimates for the WISC-R IQs and ELPs by racial-ethnic and SES groups. *School Psychology Review, 12*, 57–61.

Oplesch, M., & Genschaft, J. L. (1981). Comparison of bilingual children on the WISC-R and the Escala de Inteligencia Wechsler Para Ninos. *Psychology in the Schools, 18*, 159–163.

Osterlind, S. J. (1989). *Constructing test items*. Boston, MA: Kluwer.

Plotnik, R. (1996). *Introduction to psychology* (4th ed.). Pacific Grove, CA: Brooks/Cole.

Poteat, G. M., Wuensch, K. L., & Gregg, N. B. (1988). An investigation of differential prediction with the WISC-R. *Journal of School Psychology, 26*, 59–68.

Pugh, G. M., & Boer, D. P. (1989). An examination of culturally appropriate items for the WAIS-R Information subtest with Canadian subjects. *Journal of Psychoeducational Assessment, 7*, 131–140.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13–16.

Reschley, 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: Guilford.

Reynolds, C. R. (1980). An examination for bias in a preschool test battery across race and sex. *Journal of Educational Measurement, 17*, 137–146.

Reynolds, C. R. (1983). Regression analyses of race and sex bias in seven preschool tests. *Journal of Psychoeducational Assessment, 1*, 169–178.

Reynolds, C. R. (1998). Race bias in testing. In R. J. Corsini & A. J. Auerbach (Eds.), *Concise encyclopedia of psychology* (2nd ed., p. 740). New York: Wiley.

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

Reynolds, C. R. (in press). Why do we ignore research on bias in mental testing? *Psychology, Public Policy, and Law*.

Reynolds, C. R., & Brown, R. T. (Eds.). (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: Plenum.

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

Reynolds, C. R., & Piersel, W. C. (1983). Multiple aspects of bias on the Boehm Test of Basic Concepts (Forms A & B) for White and for Mexican-American children. *Journal of Psychoeducational Assessment, 1*, 135–142.

Reynolds, C. R., Lowe, P. A., & Saenz, A. L. (1999). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology* (3rd. ed., pp. 549–595). New York: Wiley.

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

Richardson, T. Q. (1993). Black cultural learning styles: Is it really a myth? *School Psychology Review, 22*, 562–567.

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

Rosenbach, J. H., & Mowder, B. A. (1981). Test bias: The other side of the coin. *Psychology in the Schools, 18*, 450–454.

Rotberg, I. C. (1995). Myths about test score comparisons. *Science, 270*, 446–448.

Sandoval, J., & Miille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology, 48*, 249–253.

Sandoval, J., Zimmerman, I. L., & Woo-Sam, J. M. (1983). Cultural differences on WISC-R verbal items. *Journal of School Psychology, 21*, 49–55.

Scarr, S. (1981). Implicit messages: A review of Bias in Mental Testing. *American Journal of Education, 89*, 330–338.

Scarr-Salapatek, S. (1971). Race, social class, and IQ. *Science*, 174, 1285–1295.

Sharpley, C. F., & Stone, J. M. (1985). An exploratory investigation to detect cross-cultural differences on the PPVT-R. *Psychology in the Schools, 22*, 383–386.

Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement*. New York: Guilford.

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

Silverstein, A. B., Belger, K. A., & Morita, D. N. (1982). Social class differences on the Boehm Test of Basic Concepts: Are they due to bias? *Psychology in the Schools, 19*, 431–432.

Snyderman, M., & Rothman, S. (1988). *The IQ controversy, the media and public policy*. New Brunswick, NJ: Transaction.

Spearman, C. (1927). *The abilities of man*. New York: MacMillan.

SPSSI. (1969). The SPSSI Statement. *Harvard Educational Review, 39*, 625–627.

Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence? Contemporary viewpoints on its nature and definition*. Norwood, NJ: Ablex.

Stone, B. J. (1992). Prediction of achievement by Asian-American and White children. *Journal of School Psychology, 30*, 91–99.

Stone, B. J., & Gridley, B. E. (1991). Test bias of a kindergarten screening battery: Predicting achievement for White and Native American elementary students. *School Psychology Review, 20*, 132–139.

Suzuki, L. A., & Valencia, R. R. (1997). Race-ethnicity and measured intelligence. *American Psychologist, 52*, 1103–1114.

Suzuki, L. A., Meller, P. J., & Ponterotto, J. G. (Eds.). (1996). *Handbook of multicultural assessment: Clinical, psychological, and educational applications*. San Francisco, CA: Jossey-Bass Publishers.

Taylor, R. L. (1991). Bias in cognitive assessment: Issues, implications, and future directions. *Diagnostique, 17*, 3–5.

Valencia, R. R. (1984). Reliability of the Raven Coloured Progressive Matrices for Anglo and for Mexican-American children. *Psychology in the Schools, 21*, 49–52.

Valencia, R. R., & Rankin, R. J. (1985). Evidence of context bias on the McCarthy Scales with Mexican American children: Implications for test translation and nonbiased assessment. *Journal of Educational Psychology, 77*, 197–207.

Valencia, R. R., & Rankin, R. J. (1988). Evidence of bias in predictive validity on the Kaufman Assessement Battery for Children in samples of Anglo and Mexican American children. *Psychology in the Schools, 25*, 257–263.

Valencia, R. R., Rankin, R., & Livingston, R. (1996). K-ABC content bias: Comparisons between Mexican American and White children. *Psychology in the Schools, 32,* 153–169.

Van de Vijver, F., & Poortinga, Y. H. (1994). Bias: Where psychology and methodology meet. In A. Bouvy, F. van de Vijver, P. Boski, & P. Schmitz (Eds.), *Journeys into cross-cultural psychology* (pp. 111–126). Netherlands: Swets & Zeitlinger.

Van de Vijver, F., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13,* 29–37.

Weiss, L., & 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.

Weiten, W. (1997). *Psychology: Themes and variations* (4th ed.). Belmont, CA: Brooks/Cole.

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

Williams, R. L. (1971). Abuses and misuses in testing black children. *Counseling Psychologist, 2,* 62–77.

Williams, R. L. (1974, May). Scientific racism and IQ: The silent mugging of the black community. *Psychology Today,* 32–41.

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.

Action Editor: Craig L.Frisby
Acceptance Date: June 25, 1999