HANDBOOK OF INTELLIGENCE

THEORIES, MEASUREMENTS, AND APPLICATIONS

BENJAMIN B. WOLMAN, Editor

Kaplan, R.M. (1985) The controversy related to the use of psychological tests. In B.J. Wolman (Ed) Handbook of Intelligence: Theories, Measurements, and Applications New York: Wiley-Interscience pp 465-504.

A WILEY-INTERSCIENCE PUBLICATION

JOHN WILEY & SONS

New York Chichester Brisbane Toronto Singapore

1985

ELEVEN

THE CONTROVERSY RELATED TO THE USE OF PSYCHOLOGICAL TESTS

ROBERT M. KAPLAN

San Diego State University, and University of California, San Diego
La Jolla, California

Emotional public debates about the meaning of psychological tests have become common occurrences in classrooms, courtrooms, and professional conversations. This chapter reviews the issue of test bias, which is at the heart of the controversy. The issue of test bias is so controversial that it has inspired legislation controlling the use of tests to evaluate minority group members. Other legislation and judicial decisions have forced major changes in the testing industry.

Although test bias is the unmistakable issue of the day, we should not give the impression that it is the first controversy about mental testing. Controversy has surrounded mental testing since test reports began in 1905, and the issues have been debated on and off since the 1920s (Cronbach, 1975; Haney, 1981).

WHY IS THE ISSUE CONTROVERSIAL?

A basic tenet of U.S. society is that all people are created equal. This cornerstone of political and social thought is clearly defended in the Constitution. Yet all individuals are not treated equally, and the history of social action is replete with attempts to remedy this situation. Psychological tests are among the many practices that counteract the idea that all people are the same. Tests are designed to measure differences between people, and often the differences tests measure are in desirable personal characteristics such as intelligence and aptitude. Test scores that demonstrate differences between people may suggest to some that people are not created with the same basic abilities.

The most aggravating problem is that certain ethnic groups, on the average, score differently on some psychological tests. The most controversial case concerns intelligence tests. On the average, black Americans score 15 points or one

standard deviation lower than do white Americans on standardized IQ tests. Nobody disagrees that the two distributions greatly overlap and that there are some Blacks who score as high as the highest whites. There are also some whites who score as low as the lowest blacks. Yet only about 15% to 20% of the White population score below the average black score (Jensen, 1980).

The dispute has not concerned whether these differences exist, but rather has focused on where the responsibility for the differences lies. Many have argued that the differences are due to environmental factors (Kamin, 1974; Rosenthal & Jacobson, 1968), while others have suggested that the differences are biological (Jensen, 1969, 1972; Munsinger, 1975). The environmental versus the biological debate continues to flourish and is the topic of many publications (Loehlin, Lindzey & Spuhler, 1975). This chapter will not consider the nature-nuture question. Instead, the focus will be on tests and their use.

The review in this chapter is not limited to IQ tests. The principles discussed here are also applicable to achievement and aptitude tests. As Anastasi (1980) has suggested, it is often difficult to distinguish between the constructs of aptitude, achievement, and intelligence.

PSYCHOMETRIC STUDIES OF BIAS

This section considers the following question: Are standardized tests as valid for blacks and other minority groups as they are for whites? All the types of validity must be evaluated when the issue of test bias is considered (Cole, 1981). Some psychologists argue that the tests are differentially valid for black and white people. Because the issue of differential validity is so controversial and so emotionally arousing, it has forced a careful rethinking of many issues in test validation. Differences between ethnic groups on test performance do not necessarily indicate that the test is biased. The question is whether the test has different meanings for different groups. In psychometrics, validity defines the meaning of a test.

Item Content

Many researchers also argue that intelligence or aptitude tests are affected by language skills that are inculcated as part of a white, middle-class upbringing but are foreign to inner-city children (Kagan, Moss, & Siegel, 1963; Lesser, Fifer, & Clark, 1965; Mercer, 1971; Pettigrew, 1964; Scarr-Salapatek, 1971; Woodring, 1966). As a result of being unfamiliar with the language, some children have no chance of doing well on standardized IQ tests. For example, an American child is not likely to know what a schilling is, but a British child probably does. Similarly, the American child would not be expected to know where one puts the petrol. We assume that only a British child would understand this term. Some psychologists argue that asking an inner-city child about opera is just as unfair as asking an American child about petrol. In both cases, the term is not familiar to the child (Hardy, Welcher, Mellits, & Kagan, 1976).

Flaugher (1978) considered the accusations about the bias in psychological tests and concluded that many of them are based on misunderstandings. Many people feel that a fair test is one that asks questions they can answer. By contrast, a biased test is one that does not reveal all the test taker's strengths. Flaugher argued that the purpose of aptitude and achievement tests is to determine whether a person knows certain bits of information that are drawn from large potential pools of items. The test developers are indifferent to the opportunities people have to learn the information on the tests. The meaning they eventually assign to the tests derives from correlations of the test scores with other variables.

It has been argued that the linguistic bias in standardized tests does not cause the observed differences (Clarizio, 1979a). For example, Quay (1971) administered the Stanford-Binet to 100 children in an inner-city Head Start program. Half of the children in this sample were given a version of the test that used a black dialect, while the others were given the standard version. The results demonstrated that the advantage produced by having the test in a black dialect translates into less than a one-point increase in test scores. This finding is consistent with other research findings demonstrating that black children can comprehend standard English about as well as they can comprehend nonstandard black dialect (Clarizio, 1979a; Copple & Succi, 1974). This finding does not hold for white children, who seem to be functional only in standard dialect.

Systematic studies have failed to demonstrate that biased items in well-known standardized tests are responsible for the differences between ethnic groups (Flaugher, 1978). One approach has been to allow expert judges to eliminate particular unfair items. Unexpectedly, the many attempts to purify tests using this approach have not yielded positive results. In one study 16% of the items in an elementary reading test were eliminated after experts reviewed them and labeled them as potentially biased toward the majority group. However, when the new version of the test, which had the bad items "purged", was used, the differences between the majority and the minority school populations were no smaller than they had been when the original form of the test was used (Biachini, 1976).

Another approach to the same problem is to find classes of items that are most likely to be missed by members of a particular minority group. If a test is biased against that group, there should be significant differences between the minority and nonminority groups on certain categories of items. These studies are particularly important because if they identify certain types of items that discriminate between groups, these types of items can be avoided on future tests. Again, the results have not been encouraging; studies have not been able to identify clearly categories of items that discriminate between groups (Flaugher, 1974). The studies do show that groups differ on certain items, but it has not been clear whether these are real or chance differences. When groups are compared on a large number of items, some differences will occur for chance reasons.

A different strategy is to find items that systematically show differences between ethnic groups. Then these items are eliminated, and the test is rescored. In one study, 27 items from the SAT were eliminated because they were the specific items on which ethnic groups differed. Then the test was rescored for

everyone. Although it seems as though this procedure should have eliminated the differences between groups, it actually had only slight effects because the items that differentiated the two groups tended to be the easiest items in the set. When these items were eliminated, the test was harder for everyone (Flaugher & Schrader, 1978).

There is at least some evidence that test items do not accurately portray the distribution of sexes and races in the population. Zores and Williams (1980) reviewed the WAIS, WISC-R, Stanford-Binet, and Slosson Intelligence test items for race and sex characterization and found white males shown with disproportionate frequency. Nevertheless, it has not yet been established that bias in the frequencies with which different groups are pictured in items is relevant to the issue of test bias. Various studies have failed to demonstrate that there is serious bias in item content. Most critics argue that the verbal content of test items is most objectionable because it is unfamiliar to minority groups. However, Scheuneman (1981) reviewed the problem and concluded that the verbal material in tests is usually closer to the life experiences of blacks than is the nonverbal material.

Other statistical models have been employed to evaluate item fairness. Across these different studies, with different populations and different methods of analysis, little evidence has been produced for bias in test items (Gotkin & Reynolds, 1981). However, different models may identify different items as biased. In one comparison Ironson and Sebkovial (1979) applied four different methods to analyze item bias in the National Longitudinal Study test battery. Three methods (chi-square for group differences, transformed item difficulty, and item characteristic curves) identified many of the same items as biased in evaluating 1,691 black high school seniors contrasted to 1,794 white twelfth graders. However, there was little agreement between these item evaluations and the bias items selected using a method proposed by Green and Draper (1972).

Recently, there has been debate about the effects of biased test items upon the differential validity of a test. In one theoretical example, 25% of the items on a test were presumed to be so biased that minority test takers would be expected to perform at chance level. Despite random performance, according to this simulation, there would be only slight and perhaps undetectable differences in validity coefficients for minority and majority group members (Drasgow, 1982). One year after the publication of this paper, Dobko and Kehoe (1983) reported that the result was artificial and dependent on an unusual usage of the term "test bias." Using a more general definition of test bias and biased items, they suggested that failure to find differences in validity coefficients is consistent with the belief that the tests are equally valid for members of different ethnic and racial groups.

In summary, studies have tended not to support the popular belief that items have different meanings for different groups. However, we must continue to evaluate the fairness of test items. On some occasions careful reviews of tests have identified questionable items. Many tests are carelessly constructed, and every effort should be taken to purge items that have the potential for being biased.

Criterion Validity

College administrators who use standardized test scores to forecast first-year performance are faced with difficult problems. On the average, minority applicants have lower test scores than do nonminority applicants. At the same time most universities and colleges are attempting to increase their minority enrollments. Because minority applicants are considered as a separate category, it is appropriate to ask whether the tests have differential predictive power for the two groups of applicants.

Criterion validity of a test is typically evaluated using the coefficient of correlation between the test and some criterion and by examining regression plots and scatter diagrams. If college grades are the criterion (the variable we are trying to forecast), the validity of a test such as the SAT would be represented by the correlation between the SAT and first-year college grades.

To understand criterion validity, it is often valuable to study regression lines separately for different groups. Figure 1 shows a regression line that represents each of two groups equally well. Group A appears to be performing less well than Group B on both the test (predictor) and the criterion scores. For example, the regression for Group A and for Group B have the same slope and intercept. There is little evidence for test bias in Figure 1; Group B has high scores on the test and exhibits better performance on the criterion.

Figure 2 represents a different situation. In this instance, there is a separate regression line for each group. The slopes of the two lines are the same, and that is why the two are parallel. However, the intercepts, or points at which the lines cross the vertical axis, differ. A particular test score gives one expected criterion score for regression line A and another expected criterion score for regression line B. For a test score of 8, the expected criterion score from regression line A is 6, while the expected criterion score from regression line B is 10. The

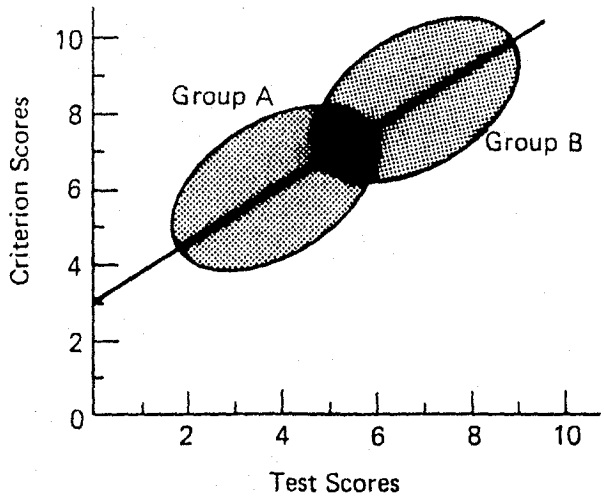

Figure 1. A single regression slope can predict performance equally well for two groups. However, the means for the two groups differ. (*Source:* From *Psychological Testing: Principles, Applications and Issues,* by R. Kaplan and D. Sacuzzo. Copyright (©) 1982 by Wadsworth, Inc. Reprinted by permission of Brooks/Cole Publishing Company, Monterey, California.)

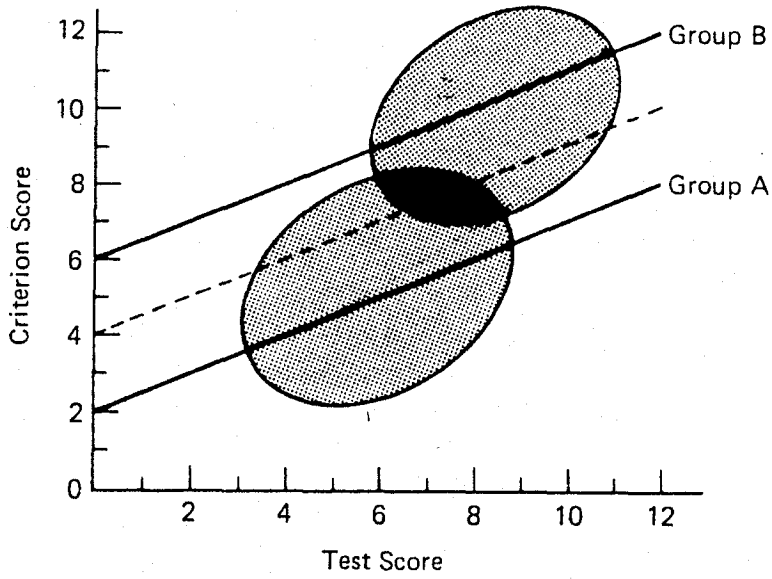

Figure 2. Regression lines with equal slopes but different intercepts. (*Source:* From *Psychological Testing: Principles, Applications and Issues,* by R. Kaplan and D. Sacuzzo. Copyright (©) 1982 by Wadsworth, Inc. Reprinted by permission of Brooks/Cole Publishing Company, Monterey, California.)

dotted line in Figure 2 is based on a combination of regression lines A and B. A test score of 8 from this combined (dotted) regression line gives an expected criterion score of 8. Thus the combined regression line actually overpredicts performance on the criterion for Group A and underpredicts performance for Group B. According to this example, the use of a single regression line produces discrimination in favor of Group A and against Group B.

Some evidence suggests that this situation is descriptive of the relationship between the SAT and college grade point average (Cleary, 1968, Kallingal, 1971; Pfeifer & Sedlacek, 1971; Temp, 1971). Each of the studies cited above showed that the relationship between college performance and SAT scores was best described by two separate regression equations. Using a combined regression equation, which is commonly the case in practice, overpredicts how well minority students will do in college and tends to underpredict the performance of majority-group students. In other words, it appears that the SAT used with a single regression line yields biased predictions, and the bias is in favor of minority groups and against majority group students.

Since the lines in Figure 2 are parallel, the slope of the lines is about the same for each group. The equal slopes suggest equal predictive validity. Most standardized intelligence, aptitude, and achievement tests do confirm the relationships shown in the figure (Reynolds, 1980; Reynolds & Nigl, 1981; Reschly & Sabers, 1979). Thus there is little evidence that tests such as the SAT predict college performance differently for different groups or that IQ tests have different correlations with achievement tests for black, white, or Hispanic children. This finding has been reported for the SAT (Temp, 1971), preschool tests (Reynolds, 1980), and IQ tests such as the WISC-R (Reschly & Sabers, 1979). Whether

separate or combined regression lines are used depends on different definitions of bias. (We return to this issue later in the chapter. The interpretation of tests for assessing different groups can be strongly influenced by personal and moral convictions.) It is worth noting that the situation shown in Figure 2 is independent of differences in mean scores. The differences in mean scores in the figure are equal to the differences between the two regression lines.

A third situation outlined by Cleary and her colleagues (Cleary, Humphreys, Kendrick, & Wesman, 1975) is shown in Figure 3. In this figure, there are two regression lines, but the lines are no longer parallel. In this situation, the coefficient for one group is different from the coefficient for the other group. In the situation presented in Figure 2, we found that each group was best represented by its own regression line. In this case, using a common regression line causes error in predicting scores for each group. However, the situation depicted in Figure 2 is not hopeless, and indeed some practitioners feel that this situation is useful because it may help increase the accuracy of predictions (Cleary, 1968). However, Figure 3 demonstrates a more hopeless situation. In this case the test is differentially valid for the two groups, meaning that the test will have an entirely different meaning for each group. Although empirical studies have rarely turned up such a case, there are some known examples of differential slopes (Mercer, 1979). An extensive discussion of differential validity is presented by Bartlett and O'Leary (1969).

ALTERNATIVE TESTS

To many American psychologists the defense of psychological tests has not been totally satisfactory. Those who do not think that the tests are fair suggest one

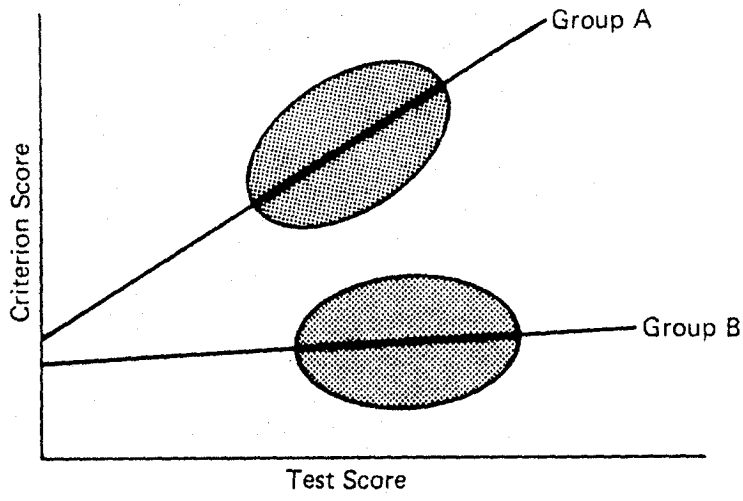

Figure 3. Regression lines with different slopes suggest that a test has different meanings for different groups. This is the most clear-cut example of test bias. (*Source:* From *Psychological Testing: Principles, Applications and Issues,* by R. Kaplan and D. Sacuzzo. Copyright (©) 1982 by Wadsworth, Inc. Reprinted by permission by Brooks/Cole Publishing Company, Monterey, California.)

of two alternatives: restrict the use of psychological tests for minority students (Williams, 1974), or develop psychological assessment strategies that suit minority children. Advocates of the first alternative have launched a legal battle to establish restrictions on the use of tests. This group emphasizes that we must try to find selection procedures that will end all discriminatory practices and protect the interests of minority-group members.

In this section various approaches to the second alternative are reviewed. In particular, five different assessment approaches are examined: the Chitling Test, the Black Intelligence Test of Cultural Homogeneity, the System of Multicultural Pluralistic Assessment, the Kaufman Assessment Battery for Children, and a Reaction time measure. Each of these approaches is different, yet they are all based on one common assumption: minority children have not had the opportunity to learn how to answer items on tests that reflect traditional, white, middle-class values.

Ignorance versus Stupidity

In a California trial about the use of testing in public schools, *Larry P. v. Wilson Riles*, the judge made an abrasive but insightful comment. Both sides in the case agreed that minority children perform more poorly on the standardized tests. The major issue debated by the witnesses was the meaning of the scores. One side argued that the scores reflect the underlying trait of intelligence. In other words, they allegedly measure how smart a child is. Witnesses for the other side suggested that the tests measure only whether the child has learned the appropriate responses needed to perform well on the test. This position claims that the tests do not measure how smart the child is, but only whether the child has been exposed to the information on the test. After hearing the testimony for the different points of view, the judge concluded that the issue was really one of ignorance versus stupidity. Although this comment appears abrasive and racist, it is quite insightful. There are two potential explanations for why some children do more poorly on standardized tests than do other children. One explanation is that they are less intelligent. In the words of the judge, this would be the stupidity explanation. The other explanation is that some children performed more poorly because they are ignorant. In other words, they simply have not learned the right responses for a particular test. If ignorance is the explanation, differences in IQ scores are less to be concerned about because they can be changed. The stupidity explanation is more damaging because it implies that the lower test scores obtained by black students are a product of some deficit that cannot be easily changed.

The term ignorance implies that differences can easily be abolished. Just as some minority children are ignorant about how to answer items that might predict success in the white, middle-class culture, some white, middle-class children could be labeled ignorant about how to succeed in the world of a ghetto child. This proposition is illustrated by the Chitling Test.

The Chitling Test

Many years ago animal psychologists talked about higher and lower animals. The higher animals were considered to be intelligent because they could do some of the same things humans can do, and the lower animals were considered to be unintelligent because they could not perform like humans. However, in 1969, a famous article by Hodos and Campbell changed the thinking of many students of animal behavior. Hodos and Campbell argued that all animals are equally intelligent for the environments in which they live. We cannot compare the intelligence of a rat with that of a cat because a rat is adapted to a rat's environment and a cat is adapted to a cat's environment. Both animals are best suited to survive in the environment they occupy.

The same insight seems not to have permeated the world of human affairs. Because of poverty and discrimination, minority and nonminority children grow up in different environments. To be successful in each of these environments requires different skills and knowledge. A psychological test may consider survival in only one of these environments, and this is usually the white, middle-class environment. Using one of these tests for impoverished children is analogous, therefore, to testing a cat on a task designed to determine how well a rat is adapted to a rat's environment.

The Chitling Test was developed by black sociologist Adrian Dove to demonstrate that there is a body of information about which the white middle class is ignorant. Dove named his effort the Dove Counterbalance General Intelligence Test, but it has become known as just the Chitling Test ("Taking the Chitling Test," 1968). A major aim in developing the Chitling Test was to show that blacks and whites have different approaches to communication.

Items on the Chitling Test ask about the definitions of a "handkerchief head," a "gas head," a "blood," "Dixie Hummingbirds," and several other items. In 1968, those who had grown up in a ghetto outperformed white, middle-class students.

However, no more than face validity has been established for the Chitling Test. No body of evidence demonstrates that the test successfully predicts performance on any important criterion. In fact, standardized tests predict performance for both minority and nonminority students, and the Chitling Test predicts performance for neither group. The Chitling Test may turn out to be a valid test of how streetwise someone is. Yet any generalizations must await validity evidence. Dove described his efforts to develop an intelligence test as "half-serious." But we have seen that the test does identify an area of content on which the races differ and blacks outperform whites.

The Black Intelligence Test of Cultural Homogeneity

To many observers, the use of intelligence tests is seen as a subtle and dangerous form of racism. Since the tests are supported by validity studies, they are given

the endorsement of scientific objectivity (Garcia, 1981). Robert Williams, a well-known black psychologist, has labeled this phenomenon scientific racism (1974). Williams views IQ and standardized achievement tests as "nothing but updated versions of the old signs down South that read 'for Whites only' (1974, p. 34).

Of particular interest to Williams and his colleagues is the assessment of survival potential with a Survival Quotient (SQ). This quotient is more important than is assessment of IQ, which only indicates the likelihood of succeeding in the white community. As a beginning, Williams developed the Black Intelligence Test of Cultural Homogeneity (BITCH), which asks respondents to define 100 vocabulary words relevant to Afro-American culture. The words came from the Afro-American Slang Dictionary and from William's personal experience interacting with black Americans. Black people obtain higher scores than did their white counterparts on the BITCH. When Williams administered the BITCH to 100 16- to 18-year-olds from each group, the average score for black subjects was 87.07 (out of 100). The mean score for the whites was significantly lower (51.07). Williams argues that traditional IQ and achievement tests are nothing more than culture-specific tests that assess how much white children know about white culture. The BITCH is also a culture-specific test, but one on which the black subjects outperform the whites.

Although the BITCH does tell us a lot about the cultural loading in intelligence and achievement tests, it has received mixed reviews. The reliability data reported by Williams show that the BITCH is quite reliable for black test takers (standard error less than 3 points on the 100-point scale) and acceptably reliable for white test takers (standard error about 6). (Conventional tests have similar reliabilities for both groups; Oakland & Feigenbaum, 1979.) However, little convincing data on the BITCH are available. Although the test manual does report some studies, the samples are small and not representative of any clearly defined population (Cronbach, 1978). Low correlations between the BITCH and California Achievement Test are reported in the manual (Williams, 1972). The difficulty is that we cannot determine whether the BITCH does predict how well a person will survive on the streets, how well he or she will do in school, in life, or in anything else. The test does assess word association, but it seems to give no information on reasoning abilities.

Further studies are needed to determine whether the BITCH does what it is supposed to do. One of the rationales for the test is that it will identify children who have been unfairly assigned to classes for the Educable Mentally Retarded (EMR) on the basis of IQ scores. In one study, Long and Anthony (1974) attempted to determine how many Black EMR children would be reclassified if they were retested with the BITCH. Among a small and limited sample of 30 Black EMR high school students from Gainesville, Florida, all the students who performed poorly on the WISC also performed below the first percentile on the BITCH. Using the BITCH served to reclassify none of the students. In another study, middle-class black seventh graders obtained higher BITCH scores than did their white middle-class counterparts. However, there were no differences between lower-class blacks and whites (Andre, 1976). These data do not

strongly support the value of the BITCH. In its present state, the BITCH can be a valuable tool for measuring white familiarity with the black community. When white teachers or administrators are sent to schools that have predominantly black enrollments, the BITCH may be used to determine how much they know about the culture. Furthermore, the BITCH may be used to assess the extent to which a black person is in touch with his or her own community and may be useful in building black pride (Milgram, 1974). As Cronbach (1978) notes, people with good abstract reasoning skills may function poorly if they are unfamiliar with the community in which they live. Similarly, people with poor reasoning skills may get along just fine if they are familiar with the community.

The System of Multicultural Pluralistic Assessment

The system of Multicultural Pluralistic Assessment (SOMPA) (Mercer, 1979), developed by sociologist Jane Mercer, offers a strong challenge to traditional views of testing. Before reviewing the SOMPA and the evaluations of it, it is instructive to review Mercer's beliefs about the social and political implications of testing.

Mercer argues that beliefs about fairness are related to the social structure. She agrees with sociologist K. Mannheim (1936) that members of the politically dominant group provide the interpretation of events within a society and they do so from their own perspective. The traditional psychometric literature on IQ tests provides a scientific rationale for the dominant group to restrict minority-group members by demonstrating that the minority-group members do not have the language and knowledge skills to perform well in a white cultural setting. The feedback given to the minority groups is not that they are ignorant about the rules of success in another culture (just as the dominant group would be in a minority culture), but that they are stupid and unlikely to succeed. Mercer emphasizes that we must take into consideration that some individuals are working from a different knowledge base.

It is not possible to give a complete description of the SOMPA here. The system is complex, and many technical issues have been raised about its validity and its applicability (Brown, 1979a, 1979b; Clarizio, 1979a, 1979b; Goodman, 1977, 1979; Mercer, 1979; Oakland, 1979).

One important philosophical assumption underlies the development of the SOMPA. This assumption is that all cultural groups have the same average potential. Any differences between cultural groups are assumed to be caused by differences in access to cultural experiences. Those who do not perform well on the tests are not well informed about the criteria for success that are usually set forth by the dominant group. However, within groups that have had the same cultural experiences, not all individuals are expected to be the same; and assessments of these differences is a better measure of ability than is assessment of differences between cultural groups.

Mercer has been concerned about the consequences of labeling a child as

mentally retarded (Mercer, 1972). She has convincingly argued that many children are incorrectly identified as retarded and that they suffer severely as a result of this inaccurate branding. In particular, she is distressed that classes for EMR students contain a disproportionate number of minority children. Mercer maintains that some minority students score low on the traditional tests because they are merely ignorant about the ways of the dominant culture, but they are not mentally retarded. Students may also be misclassified due to medical problems. Thus a fair system of evaluation must include medical assessment. It must also include the assessment of children relative to other children who have had similar life experiences. The basic point of divergence between the SOMPA and earlier approaches to assessment is the SOMPA attempts to integrate three different approaches to assessment: medical, social, and pluralistic.

SOMPA System. One of the most consistent findings in the field of public health is that members of low-income groups have more health problems than do those who are economically better off. The medical component of the SOMPA system asks: "Is the child an intact organism?" (Mercer, 1979, p. 92). The rationale for this portion is that medical problems can interfere with a child's performance on mental measures and in school.

The social system component attempts to determine whether a child is functioning at a level that would be expected by social norms. For example, does the child do what is expected by family members, peer groups, or the community? Mercer feels that test users and developers typically adopt only a social-system orientation. For example, if a test predicts who will do well in school, it is forecasting behavior that is expected by the dominant social system. However, Mercer emphasizes that the social-system approach is a narrow one because only the dominant group in society defines the criteria for success (Reschly, 1981).

The pluralistic component of the SOMPA recognizes that different subcultures are associated with different life experiences. Only within these subgroups do individuals have common experiences. Thus tests should assess individuals against others in the same subculture. It is important to recognize the distinction between the criteria for defining deviance in the pluralistic model and in the social-system mode. The social-system model used the norms of society as the criteria, while the pluralistic model uses the norms within a particular group.

The SOMPA attempts to assess children relative to each of these models. The medical portion of the SOMPA package includes physical measures such as visual tests, tests of hearing, and tests of motor functioning. The social-system portion is similar to most assessment procedures. The entire WISC-R is given and evaluated according to the regular criteria. Finally, the pluralistic portion also uses WISC-R scores but evaluates them against those for groups that have similar social and cultural backgrounds. In other words, the WISC-R scores are adjusted for socioeconomic background. These adjusted scores are known as estimated learning potentials (ELP).

The major dispute between Mercer and her many critics concerns the validity of the SOMPA. Mercer (1979) points out that a test itself is not valid or invalid

but the inferences that are made on the basis of the test scores are. She insists that the ELPs cannot be validated in the same way as are other test scores. (Validating a test by predicting who will do well in school is appropriate only for the social-system model.) Mercer argues that the criteria for evaluating the ELP must be different. She states that the appropriate validity criteria for ELPs should be the percentage of variance in WISC-R scores that is accounted for by sociocultural variables. Many SOMPA critics (Brown, 1979a; Clarizio, 1979b; Goodman, 1979; Oakland, 1979), however, feel that a test should always be validated by demonstrating that it predicts performance. The correlation between ELPs and school achievement is around .40, while the correlation between the WISC-R and school achievement is around .60 (Oakland, 1979). Thus ELPs are a poorer predictor than WISC-R scores of school success. Mercer refutes these critics by arguing that the test is not designed to identify which children will do well in school. Its purpose is to determine which children are mentally retarded. Yet it gives minority students additional IQ points to compensate for their impoverished backgrounds. In one example, Sattler (1982) showed that the system can boost a child with a full-scale WISC-R score in the 2nd percentile all the way up to the 70th percentile. This can be done only by comparing children with others who have had the same life experiences.

Accepting Mercer's argument may produce a quota system for EMR classes. Using ELPs should make the proportions of ethnic groups in EMR classes more representative than they now are. Because several states have adopted the SOMPA, we may soon be able to determine the ultimate effect of the system. There is no question that it will identify fewer minority children as EMR students. This may produce cost reductions because the costs of educating EMR students are higher than average. Only time will tell whether children no longer considered EMR students will benefit. Mercer's (1972) research suggests that a big part of the battle is just getting more children labeled as normal. Her critics retaliate by suggesting that the effects of labeling are weak and inconsequential (Thorndike, 1968). They argue that no matter what these children are called they will need some special help in school.

The Kaufman Assessment Battery for Children (KABC)

In 1983, Kaufman and Kaufman introduced a new approach to the assessment of intellectual abilities in children. Their tests, known as the Kaufman Assessment Battery for Children (K-ABC), separates two fundamental components of human information processing: simultaneous processing and sequential processing. Similar distinctions have been made by several cognitive psychologists with only slight variations in the label. For instance, the Kaufman's Sequential-Simultaneous distinction is quite similar to the Successive-Simultaneous distinction (Das, Kirby and Járman, 1975, 1979; Luria, 1966), the Analytic-Holistic Distinction (Ornstein, 1972) or the Serial-Parallel/Sequential-Multiple Distinction (Neisser, 1967). As Kaufman (1983) argues, there has been a strong concensus that there is a basic dichotomy in types of human information processing. However, there

has been less concensus about the neuroanatomical site responsible for this distinction. Some attributed it to a temporal/occipital-parietal distinction while others describe it as left brain/right brain difference (Kaufman and Kaufman, 1983).

The K-ABC is an individually administered intelligence test that was standardized on a nationwide (American) sample of normal and exceptional children ranging in age from 2½ to 12½ years. In each age range, the test measures sequential and simultaneous information processing. In addition, it has a separate section for achievement. One of the features of the K-ABC is that it proportedly does not have the same racial bias that characterizes most other IQ tests. Data for the WISC-R Standardization program show the mean score of white children (ages 6 through 16) to be 102.3, while the mean for black children in the same age range is 86.4 (Kaufman and Doppelt, 1976). A separate study on WISC-R data showed the mean for a sample of Hispanic children to be 91.9 (Mercer, 1979).

Mean scores for the K-ABC are considerably closer together for black and white students. For white students, K-ABC scores are nearly equivalent to the WISC-R scores (Mean = 102.0). For the 807 black students in the standardization sample, the mean was 95.0, while it was 98.9 for the 106 Hispanic students in the standardization program (Kaufman and Kaufman, 1983). For the sequential processing portion of the K-ABC, black, white, and Hispanics in the 2½- through 12½-age range performed at near equivalent levels. In addition, Hispanic students performed at near equivalent levels to the white standardization sample.

Although predictive validity data are not presented separately for race or ethnic groups, some evidence suggests that the concurrent validity of the K-ABC is comparable for different groups. For example, the correlation between the K-ABC mental processing composite score and the Woodcock Reading Mastery Test was .60 for the white sample in one validation study. For the black and Mexican-American subsamples, these correlations were .56 and .70, respectively. In another study, the K-ABC correlated slightly more highly with the KeyMath Diagnostic Arithmetic Test for black and Hispanic samples than it did for a white sample (Kaufman and Kaufman, 1983). Although a majority of the 43 validity studies reported in the manual had at least some minority subjects, these are the only studies that reported validity data separately for different ethnic or racial groups. In summary, the K-ABC appears to be a promising approach. Mean differences between racial groups are smaller than they are for other intelligence tests. Yet the K-ABC appears to have firm roots in empirical psychology and to have a substantial record of reliability and validity. Further validity data will be required to confirm or disconfirm these encouraging observations.

Factor Theory and IQ Differences

In 1927, Charles Spearman presented a discussion of racial differences in intelligence. He argued that the amount by which groups differed was not consistent

across different mental tasks. However, he suggested that there was a *g* factor or general intelligence factor that was common to many mental tasks. The *g* factor is widely discussed in the psychological literature and the concept has endured for more than a half century. The *g* factor can be obtained from a factor analysis of mental tasks. It is the primary factor representing tests of verbal, numerical, spatial, and other general mental abilities.

The degree to which black and white subjects differ in mental test scores differs across studies. Jensen (1983) argued that these different results can be explained by the tasks used in the different studies. He suggested that tasks that loaded highly on the *g* factor are more likely to show the differences than will tasks with low *g* loading. Furthermore, he suggested that scores on a reaction time task are highly correlated with the *g* factor. Black and white subjects differ in their mean performance on these reaction time tests. Nevertheless, the tasks do not require language and should be no more familiar to one group than to the other.

In Jensen's most recent work, he has been careful not to offer a causal mechanism for these differences. Yet many of his critics assume that Jensen's view of intelligence regards abilities as fixed and unchangable over the course of time. Robert Sternberg (cited in Cordes, 1983) disagrees that intelligence tests measure a fixed trait. Instead, he suggests that they measure cognitive processes that may be altered through experience. Jones (1983) and others have demonstrated that Black-White differences in standardized achievement scores narrowed between 1971 and 1980. They noted that math scores tended to diverge during the course of time. Yet multiple regression analysis demonstrated that math scores are well predicted from the number of algebra and geometry courses a student has completed. Black children took fewer of these courses and thus did more poorly on the math tests. It was argued that the difference in test scores could be irradicated by providing more mathematics training for black youth.

In summary, it was suggested that cognitive abilities measured by IQ and achievement tests can be modified by educational experiences. It is these complex abilities that load highly on the *g* factor. We expect research to shed more light on this debate in the years to come.

Ethical Concerns and the Definition of Test Bias

It is difficult to define the term *test bias* since different authors have different views (Cole, 1981; Darlington, 1978; Flaugher, 1978; Hunter & Schmidt, 1976). These different definitions represent commitments to underlying ethical viewpoints about the way various groups ought to be treated. Hunter and Schmidt (1976) identify three ethical positions that set the tone for much of the debate: unqualified individualism, the use of quotas, and qualified individualism. All these positions are concerned with the use of tests to select people either for jobs or for training programs (including college).

Supporters of unqualified individualism would use tests to select the most qualified individuals they could find. In this case, users of tests would be in-

different to the race or sex of applicants. The goal would be to predict those who would be expected to perform best on the job or in school. According to this viewpoint, a test is fair if it finds the best candidates for the job or for admission to school. If race or sex were a valid predictor of performance beyond the information in the test, the unqualified individualist would see nothing wrong with considering this information in the selection process.

A quite different ethical approach to selection is to use quotas. Quota systems explicitly recognize race and sex differences. If the population of a state is 20% black, then supporters of a quota system might argue that 20% of the new medical students in the state-supported medical school should also be black. Selection procedures are regarded as biased if the actual percentage of applicants admitted is different from the percentage in the population; each group should have a fair share of the representation (Gordon & Terrell, 1981). This fair-share selection process gives less emphasis than the testing process to how well people in the different groups are expected to do once they are selected (Darlington, 1971; Hunter & Schmidt, 1976; S. Thorndike, 1971).

The final moral position considered by Hunter and Schmidt might be viewed as a compromise between unqualified individualism and a quota system. Qualified individualism, like unqualified individualism, embraces the notion that the best-qualified persons should be the ones selected. But unqualified individualists also take information about race, sex, and religion into consideration if it helps predict performance on the criterion. Not to do so results in underprediction of performance for one group and overprediction of performance for another group. Qualified individualists, however, recognize that, although failing to include group characteristics (race, sex, and religion) may lead to differential accuracy in prediction, this differential prediction may counteract known effects of discrimination. It may, for example, lead to underprediction of the performance of the majority group and overprediction of the performance of the minority group. The qualified individualist may choose not to include information about personal characteristics in selection because ignoring this information may serve the interest of minority-group members.

Each of these ethical positions can be related to a particular statistical definition of test bias, and we now turn to these definitions. Table 1 shows several different models of test bias based on different definitions of fairness. All these models are based on different definitions of fairness. All these models are based on regression lines as we discussed above. The models discussed in Table 1 are relevant to tests that are used for selection purposes, including job-placement tests and tests used to select students for college or for advanced-degree programs.

The straight regression approach described in Table 1 (see also Cleary, 1968) represents the unqualified individualism position. The result of this approach is that a large number of majority-group members may be selected. In other words, this approach maintains that an employer or a school should be absolutely color and gender blind. The reason for considering ethnicity or sex is to improve prediction of future performance. This approach has been favored by business

Table 1. Different Models of Test Fairness

Model	Reference	Use of Regression	Rationale	Effect on Minority Selection	Effect on Average Criterion Performance
Regression	Cleary (1968)	Separate regression lines are used for different groups. Those with the highest predicted criterion scores are selected.	This is fair because those with the highest estimated level of success are selected.	Few minority-group members selected.	High.
Constant Ratio	R. L. Thorndike (1971)	Points equal to about half of the average difference between the groups are added to the test scores of the group with the lower score. Then a single regression line is used, and those with the highest predicted scores are selected.	This is fair because it better reflects the potential of the lower-scoring group.	Some increase in the numbr of minority-group members selected.	Somewhat lower.
Cole/Darlington	Cole (1973); Darlington (1971, 1978)	Separate regression equations are used for each group, and	This is fair because it selects more potentially successful	Larger increase in the number of minority-group mem-	Lower.

Table 1. (*Continued*)

Model	Reference	Use of Regression	Rationale	Effect on Minority Selection	Effect on Average Criterion Performance
Cole/Darlington		points are added to scores of those from the lower group to assure that those with the same criterion score have the same predictor score.	persons from the lower group.	bers selected.	
Quota	Dunnette and Borman (1979)	The proportion of persons to be selected from each group is predetermined. Separate regression equations are used to select those persons from each group who are expected to perform highest on the criterion.	This is fair because members of different subgroups are selected based on their proportions in the community.	Best representation of minority groups.	About the same as for the Cole/Darlington model.

Source. Kaplan and Saccuzzo (1982, p. 456).

Note. Based on Dunnette and Borman (1979).

because it ensures the highest rate of productivity among the employees who are selected by the procedure.

At the other extreme is the quota system. To achieve fair-share representation, separate selection procedures are developed. One procedure, for example, is used to select the best available black applicants, and another procedure is used to select the best available nonblack applicants. If a community has 42% black residents, the first procedure would be used to select 42% of the employees and the other procedure would be used to select the other 58%.

The difficulty with the quota system is that it may lead to greater rates of failure among some groups. Suppose, for example, that a test had been devised to select telephone operators and that the test did indeed predict who would succeed on the job. However, the test selected 70% women and only 30% men. The quota system would encourage the use of separate cutoff scores so that the proportion of men selected would approach 50%. But, because the women scored higher on the average, they would perform better on the job, resulting in a higher rate of failure among men. Thus, although quota systems often aid in increasing the selection of underrepresented groups, they also make it likely that the underrepresented groups will experience failure.

Table 1 shows two other models (Cole, 1973; Darlington, 1971; Thorndike, 1971). These models represent compromises between the quota and the unqualified-individualism points of view. In each of these cases, there is an attempt to select the most qualified people, yet there is some adjustment for being from a minority group. When people from two different groups have the same test score, these procedures give a slight edge to the person from the lower group and places the person from the higher group to a slight disadvantage. Although these approaches have been attacked for being based on faulty logic (Hunter & Schmidt, 1976, 1978), plausable defenses have been offered. The effect of these procedures is to increase the number of people selected from underrepresented groups. However, these procedures also result in lower expected performance scores on the criterion.

Which of these approaches is right and which is wrong? That is a value decision that embraces different philosophical beliefs about fairness.

LEGAL CONTROVERSIES

The U.S. government has attempted to establish clear standards for the use of psychological tests. Regulation of tests comes in many forms, including executive orders, laws created by legislative bodies, and actions by the courts. The most important legal development was the passage of the 1964 Civil Rights Act. Title VII of this act created the Equal Employment Opportunity Commission (EEOC). The EEOC in 1970 published guidelines for employee-selection procedures. In 1978, it released a new document entitled, "Uniform Guidelines on Employee Selection Procedure." These are the major guidelines for the use of psychological tests in education and in industry.

The 1978 guidelines are stricter, more condensed, and less ambiguous about the allowable use of psychological test scores than were the 1970 guidelines. The original act clearly prohibited discrimination in employment on the basis of race, color, religion, sex, or national origin. However, the 1978 guidelines made clear that any screening procedure, including the use of psychological tests, may be viewed as having adverse impact if it systematically rejects substantially higher proportions of minority than nonminority applicants. When any selection procedure does so, the employer must demonstrate that the procedure has documented validity. However, the guidelines are specific about the acceptable criteria for the use of a test. The guidelines have been adopted by a variety of federal agencies including the Civil Service Commission, the Department of Justice, the Department of Labor, and the Department of the Treasury. The Office of Federal Contract Compliance has the direct power to cancel government contracts held by employers who do not comply with these guidelines. In the next few years it is almost certain that these guidelines will provide the basis for law suits filed by both minority and nonminority job applicants who feel they have been mistreated in their employment pursuits.

It is worth noting that the guidelines are used only for cases in which adverse impact is suspected. When adverse impact is not suspected, organizations are under little pressure to use valid selection procedures (McCormick & Ilgen, 1980). As Guion (1976) observes, "organizations have the right to even be fairly stupid in their employment practices as long as they are stupid fairly" (p. 811).

Specific Laws

Other regulatory schemes attempt to control the use of tests. Recently, Truth in Testing Laws have been passed in two states (New York and California) and similar bills have been introduced in several other states and at the federal level.

The New York Truth in Testing Law is one of the most controversial measures ever to hit the testing field. The New York law was motivated by an extensive investigation of the Educational Testing Service (ETS) by the New York Public Interest Research Group (NYPIRG). Other testing companies are affected by the law, but the New York law was written with ETS specifically in mind.

ETS was created by the College Entrance Examination Board, the American Council on Education, and the Carnegie Foundation in 1948. Its original and best-known mission was to create and administer aptitude tests such as the SAT. By 1979, ETS was responsible for more than 300 testing programs, including the Graduate Management Admission Test (GMAT), the Graduate Record Examination (GRE), the Multi-State Bar Exam, and the Law School Admission Test (LSAT). The assets of the company exceeded $25 million, and its gross yearly income exceeded $80 million.

NYPIRG seemed upset by the wealth and success of ETS, yet what bothered NYPIRG more was the power ETS has. Each year several million people take tests designed and administered by ETS, and the results of these tests have pronounced effects on their lives (Brill, 1973; Kiersh, 1979; Levy, 1979). Many

educational programs take the scores seriously. Students scoring poorly on the LSAT, for example, may be denied entrance to law school, and this rejection may eventually affect many important aspects of their lives. Higher scores may have resulted in a higher income, more occupational status, and greater self-esteem for them.

On investigation, NYPIRG became dissatisfied with the available information on test validity, the calculation of test scores, and the financial accounting of ETS. The Truth in Testing Law responds to these objections by requiring testing companies to (1) disclose all studies on the validity of a test, (2) provide a complete disclosure to students about what scores mean and how they were calculated, and (3) on request by a student, provide a copy of the test questions, the correct answers, and the student's answers.

The first two portions are essentially noncontroversial. The test developers argue that they do disclose all pertinent information on validity, and they do release many public documents highlighting the strengths and weaknesses of their tests. Furthermore, ETS strongly encourages institutions using their tests to perform local validity studies. Any of these studies can be published in scholarly journals with no interference from ETS. However, NYPIRG provided some evidence that ETS and other testing companies have files of secret data that they do not make public because these data may reflect poorly on the product. The second aspect of the law was included because ETS sometimes reports index scores without telling students how the index was calculated and the exact index value being reported.

The controversial third portion of the law may turn out to seriously decrease the value of testing programs. Requiring that the test questions be returned to students means that the same questions cannot be used in future versions of the test. Several problems are expected to result from this policy. First, it decreases the validity of the test. With the items constantly changing, the test essentially becomes a new test each time the items change. As a result, it is impossible to accumulate a record of construct validity.

Second, it is difficult to equate scores across years. For example, a graduate school must often consider students who took the GRE in different years. If the test itself is different each year, it is difficult to compare the score of students who took the test at different times. Although the bill eventually adopted in New York did allow testing companies to keep some of the items secret for equating purposes, this practice falls short of being a satisfactory solution. Equating can be accomplished, but it may be difficult without increasing the chances of error.

Third, the most debated problem associated with the disclosure of test items is that it will greatly increase the costs to ETS and other testing companies. It has been estimated that test construction costs range from $50,000 to $165,000 (APA, Committee on Psychological Tests and Assessment, 1983). ETS will probably not absorb these inflated costs but will pass them on to the consumer. Just how high the cost of taking a test will go is a matter of conjecture. Experts within the testing industry warn that the costs could be more than twice what

they were before the law was passed. NYPIRG doubts that the cost of writing new items will have any substantial impact on the cost of taking the test. Only 5% of the students' fees for taking a test go to question development, while 22% to 27% go to company profit. Backers of the law feel there should be only minimal increases in fees and that ETS, as a nonprofit and tax-exempt institution, should take the cost of writing new items out of its substantial profits.

One immediate impact of the New York Truth in Testing Law was that it stimulated other similar proposals. A similar law was passed in California and bills were introduced in several other states. The Educational Testing Act of 1979 was offered to the U.S. House of Representatives in July 1979. The Educational Testing Act, which was proposed by Representatives Ted Weiss, Shirley Chisholm, and George Miller, was essentially the same as the New York Truth in Testing Law; the major provisions of the bills were almost identical. In effect, the Educational Testing Act of 1979 attempted to make the New York law a federal law. However, the federal bill was not enacted into law.

There is no question that the truth-in-testing bills were introduced by sincere and well-intentioned legislators. However, the laws are disturbing for two reasons. First, they politicize a process that has in the past been primarily academic. The issues in the debate were not presented in a scholarly fashion. Instead, they were presented (on both sides) in an adversarial manner. The debate thus got out of the hands of psychologists who have the training to interpret some of the complex technical issues. For example, in his testimony before a subcommittee of the House of Education and Labor Committee, Representative Weiss made many references to the bias in the tests. His major argument was that there are mean differences between different ethnic groups in test scores, yet mean difference is not usually considered evidence for test bias.

ETS does make booklets available to the public that present information on the scoring system, the validity, the reliability, and the standard error of measurement for each of their tests. People with no background in testing probably will not comprehend all this information, and the authors of the bills fail to recognize that the proper use of tests and test results is a technical problem that requires technical training in advanced courses such as psychological testing. For instance, we do not expect people to be able to practice medicine without the technical training given in medical school.

Second, we must consider the ultimate impact of the truth-in-testing legislation. One side argues that the new laws will make for a fairer and more honest testing industry. The other argues that students will now have to pay a higher price for a poorer product. If the requirement of test-item disclosure results in lower validity of the tests (which it most likely will), there will be greater error in selecting students than now exists. In other words, selection for admissions may become more random.

By summer of 1983, the Committee on Psychological Tests and Assessment of the American Psychological Association issued a formal statement on test-item disclosure legislation. The statement suggested that proposals for truth-in-testing legislation be postponed until the impact of the bills passed in California

Table 2. Summary of American Psychological Association Statement on Testing Legislation

1. We recommend a "wait and see" period prior to enacting further legislation. The legislation enacted in New York and California has created a situation that can be viewed as a naturally occurring field experiment. A few years of studying this "experiment" is warranted before other legislation is passed or rejected.
2. We strongly support provisions encouraging the dissemination of information about test content, test purpose, validity, reliability, and interpretation of test scores. Test takers should have access to their individual results and interpretative information, especially where such test results are used for educational or employment decisions.
3. We oppose total disclosure of items from low volume tests or tests where item domains are finite (for instance measuring specific content areas).
4. We oppose disclosure of tests where interpretation is dependent upon a long history of research (for example, extensive norming); disclosure would result in loss of valid interpretation. This is particularly true of interest and personality measures.
5. In disclosure cases involving large volume tests testing a broad domain, we recommend at a minimum that only items used to determine test performance be disclosed. Pretesting and equating items should be protected from disclosure.
6. Where disclosure is deemed desirable, we encourage examination of alternative methods of conveying this information to test takers, such as partial disclosure (disclosure of one test after several administrations) or making sample tests available for perusal at a secure location.
7. We urge that any personnel selection or licensing procedure in use should be subject to the same scrutiny as tests, provided that sample size is sufficient for meaningful statistical analyses.

Source. American Psychological Association Statement on Test Item Disclosure Legislation (August, 1983).

and New York could be evaluated. The committee's conclusions and recommendations are shown in Table 2.

Some Major Lawsuits

There have already been many lawsuits concerning the use of psychological tests, and the number can be expected to increase dramatically in the years to come. Some of the most important of these lawsuits are discussed in this chapter. It is important to realize that each of these cases was complex and involved considerably more evidence than can be reviewed here.

Early Desegregation Cases. The fourteenth Amendment requires that all citizens be granted the equal protection of the laws. At the end of the nineteenth century, it was being argued that segregated schools did not offer such protection. In the famous 1896 case of *Plessy v. Ferguson,** the Supreme Court ruled that

*163 U.S. 537 (1896).

schools could remain segregated, but that the quality of the schools must be equal. This was the much acclaimed separate but equal ruling.

Perhaps the most influential ruling in the history of American public school education came in the case of *Brown v. Board of Education** in 1954. In the Brown case, the Supreme Court overturned the *Plessy v. Ferguson* decision and ruled that the schools must provide nonsegregated facilities for black and white students. In its opinion the court raised several issues that would eventually affect the use of psychological tests.

The most important pronouncement of Brown was that segregation was a denial of equal protection. In coming to its decision, the court made extensive use of testimony by psychologists. This testimony suggested that black children could be made to feel inferior if the school system kept the two races separate.

The story of the Brown case is well known, but what is less often discussed is the ugly history that followed. Many school districts did not want to desegregate, and the battle over busing and other mechanisms for desegregation continues today in many areas. Many of the current arguments against desegregation are based on fear of children leaving their own neighborhoods or on the stress on children who must endure long bus rides. The early resistance to the Brown decision was more clearly linked to the racist belief of Black inferiority.

Stell v. Savannah-Chatham County Board of Education.† The most significant racist court case occurred when legal action was taken to desegregate the school system of Savannah, Georgia, on behalf of a group of black children. The conflict began when attorneys for two white children intervened. They argued that they were not opposed to desegregating on the basis of race but that black children did not have the ability to be in the same classrooms as Whites. Testimony from psychologists indicated that the median IQ score for black children was 81 while that for white children was 101. Because there was such a large difference in this trait (which was assumed to be genetic), the attorneys argued that it could be to the mutual disadvantage of both groups to congregate them in the same schools. Doing so might create even greater feelings of inferiority among black children and might create frustration that would eventually result in antisocial behavior.

The court essentially agreed with this testimony and ruled that the district should not desegregate. The judge's opinion reflected his view of what was in the best interest of all of the children. Later, this decision was reversed by Judge Griffin Bell of the U.S. Court of Appeal for the Fifth Circuit. In doing so, the court used the precedent set forth by Brown as the reason for requiring the Savannah district to desegregate. It is important to note that the validity of the

*347 U.S. 483 (1954), 349 U.S. (1955).

†220 F. Supp. 667, 668(S.D. Ga. 1963, rev'd 333 F.2d 55(5th Cir. 1964 cert. denied, 379 U.S. 933 (1964).

test scores, which were the primary evidence, was never discussed (Bersoff, 1979, 1981).

Hobson v. Hansen.* Stell was just one of many cases that attempted to resist the order set forth in the famous Brown desegregation case. Like Stell, many of these cases introduced test scores as evidence that black children were genetically incapable of learning or being educated in the same classrooms as white children. The courts routinely accepted this evidence. Given the current controversy regarding the use of psychological tests, it is remarkable that several years passed before the validity of the test scores became an issue.

The first major case to examine the validity of psychological tests was *Hobson v. Hansen.* The Hobson case is relevant to many of the current lawsuits. Unlike the early desegregation cases, it did not deal with sending black and white children to different schools. Instead, it concerned the placement of children once they arrived at a school. Although the courts had been consistent in requiring schools to desegregate, they tend to take a hands off approach with regard to placement of students in tracks once they arrived at their desegregated schools.

The Hobson case contested the use of group standardized ability tests to place students in different learning tracks. Julius W. Hobson was the father of two black children placed in a basic track by the District of Columbia School District. Carl F. Hansen was the superintendent for the district. Within the district, children were placed in honors, regular, general, and basic tracks on the basis of group ability tests. The honors track was designed to prepare children for college, while the basic track focused on skills and preparation for blue-collar jobs. Placement in the basic track makes it essentially impossible to prepare for a high income/high prestige profession.

The rub in Hobson was that racial groups were not equally represented among those assigned to the basic track. In effect, the tracking system served to racially segregate groups by placing black children in the basic track and white children in the other tracks. Psychological tests were the primary mechanism used to justify this separation. The Hobson case was decided in 1967 by Judge Skelly Wright of the federal district court of Washington, D.C. Just two years before the decision, the Supreme Court had ruled that a group is not denied equal protection by "mere classification" (Bersoff, 1979). Nevertheless, Judge Wright ruled against the use of the tracking system when based on group ability tests. After extensive expert testimony on the validity of the tests for minority children, the judge concluded that the tests discriminated against them. An interesting aspect of the opinion was that it claimed that grouping would be permissible if it were based on innate ability. The judge asserted that ability test scores were influenced by cultural experiences, and the dominant cultural group had an

*269 F. Supp. 401 (D.D.C. 1967).

unfair advantage on the tests and thereby gained admission to the tracks that provided the best preparation for high income/high prestige jobs.

Diana v. State Board of Education.* The decision in *Hobson v. Hansen* opened the door for a thorough examination of the use of standardized tests for the placement of students in EMR tracks. The case of Diana has particular implications for the use of standardized tests for bilingual children. Diana was one of nine Mexican-American elementary school children who had been placed in EMR classes on the basis of the WISC or Stanford-Binet. These nine children represented a class of bilingual children. They brought a class action suit against the California State Board of Education, contending that the use of standardized IQ tests for placement in EMR classes denied equal protection because the tests were standardized only for whites and had been administered by a non-Spanish-speaking psychometrist. Although only 18% of the children in Diana's school district had Spanish surnames, this group made up nearly one-third of the enrollment in EMR classes.

When originally tested in English, Diana achieved an IQ score of only 30. However, when retested in Spanish and English, her IQ was 79, which was high enough to keep her out of the EMR classes in her school district. Seven of the other eight plaintiffs also achieved scores high enough on retesting in Spanish to be taken out of the EMR classes.

When faced with this evidence, the California State Board of Education decided not to take the case to court. Instead, they adopted special provisions for the testing of Mexican-American and Chinese-American children. These provisions included the following:

1. If English was not the primary language, the children would be tested in their primary language.
2. Questions based on certain vocabulary and information that the children could not be expected to know would be eliminated.
3. The Mexican-American and Chinese-American children who had been assigned to EMR classes would be reevaluated with tests that used their primary language and nonverbal items.
4. New tests would be developed by the state that reflected Mexican-American culture and that were normed for Mexican-American children (Bersoff, 1979).

Later studies confirmed that bilingual children do score higher when tested in their primary language (Bergan & Parra, 1979).

The combination of the judgment in Hobson and the change in policy brought about by Diana forced many to question seriously the use of IQ tests for the assignment of children to EMR classes. However, these decisions were quite

*401 U.S. 424(a)(1971).

specific to the circumstances in the particular cases. Hobson dealt with group tests but did not discuss individual tests. However, individual tests are used more often than group tests to make final decisions for EMR placement. The ruling in Diana was limited strictly to bilingual children. These two cases were thus not relevant to black children placed in EMR classes on the basis of individual IQ tests. This specific area was left for the most important court battle of them all—*Larry P. v. Wilson Riles.*

*Larry P. v. Wilson Riles**. In October 1979, Judge Robert Peckman of the Federal District Court for the Northern District of California handed down an opinion that declared that "The use of IQ tests which had a disproportionate effect on Black children violated the Rehabilitation Act, the Education for All Handicapped Children Act, Title VI, and the 14th Amendment when used to place children in EMR classes." Attorneys for Larry P., one of six black elementary school students who were assigned to EMR classes on the basis of IQ test results, had argued that the use of standardized IQ tests to place black children in EMR classes violated both the California constitution and the equal protection clause of the fourteenth Amendment (Opton, 1979), as well as those laws mentioned above.

The court first ruled in the case of Larry P. in 1972. It found that the school district incorrectly labeled Larry as EMR and violated his right to equal educational opportunity. As a result, a preliminary injunction was issued that prohibited that particular school district from using IQ tests for EMR placement decisions. Later, the California Department of Education called for a temporary moratorium on IQ testing until another court opinion on the validity of the tests could be obtained (Opton, 1979). The Larry P. case came before the same court that had issued the preliminary injunction in order to obtain a ruling on test validity for black children.

During the trial, both sides geared up for a particularly intense battle. Wilson Riles was the black superintendent of public instruction in California; he had instituted many significant reforms that benefited minority children. Thus it was particularly awkward to have a nationally recognized spokesperson for progressive programs named as the defendant in an allegedly racist scheme.

In defense of the use of tests, Riles and the state called many nationally recognized experts on IQ tests, including Lloyd Humphreys, Jerome Sattler, Robert Thorndike, Nadine Lambert, and Robert Gordon. These witnesses presented rather extensive evidence that IQ tests, particularly the Stanford-Binet and the WISC (which were used to test Larry and others), were not biased against blacks. Although the tests had not originally been normed for black populations, studies had demonstrated that they were equally valid for use with black and white children. (Many of the arguments supporting the use of tests for all races were summarized earlier.) If the tests were not biased, then why

*442 U.S. 405 (1975).

did Larry and the others receive higher scores when they were retested by black psychologists? The defense argued that the black psychologists did not follow standard testing procedures and that IQ test scores are not changed when standardized procedures are followed.

Statements from special-education teachers were also presented. The teachers argued that the children involved in the case could not cope with the standard curriculum and that they required the special tutoring available in the EMR classes. The children had not been learning in regular classes, and the schools investigated cases in which there was doubt about the placement. For all these children, the assignment to EMR classes was deemed appropriate (Sattler, 1979).

The Larry P. side of the case also had its share of distinguished experts, including George Albee, Leon Kamin, and Jane Mercer. The arguments for Larry were varied. His lawyers argued that all humans are born with equal capacity and that any test that assigns disproportionate numbers of children from one race to an EMR category is racist and discriminatory. The witnesses testified that dominant social groups had historically used devices such as IQ tests to discriminate against less powerful social groups and that the school district had intentionally discriminated against black children by using unvalidated IQ tests. Specifically, the tests were used to keep blacks in dead-end classes for the mentally retarded, in which they would not get the training they needed to move up in the social strata. Furthermore, the plaintiffs suggested that labeling someone as EMR has devastating social consequences. Children who are labeled as EMR lose confidence and self-esteem (Mercer, 1973), and eventually the label becomes a self-fulfilling prophecy (Rosenthal & Jacobson, 1968). In other words, labeling a child as mentally retarded may cause the child to behave as though mentally retarded.

The judge was clearly persuaded more by the plaintiffs than by the defense. He declared that the tests "are racially and culturally biased, have a discriminatory impact on black children, and have not been validated for the purpose of (consigning) black children into educationally dead-end, isolated, and stigmatizing classes." Furthermore, the judge stated that the Department of Education had "desired to perpetuate the segregation of minorities in inferior, dead-end, and stigmatizing classes for the retarded."

The effect of the ruling, was a discontinuance of IQ testing to place black children in EMR classes. The decision immediately affected all black California school children who had been labeled as EMR. More than 6000 of these children must be reassessed in some other manner.

There are strong differences of opinion about the meaning of the Larry P. decision. Harold Dent, one of the black psychologists who had retested Larry P. and the other children, hailed the decision as a victory for black children:

> For more than 60 years psychologists have used tests primarily to justify the majorities desire to "track" minorities into inferior education and dead-end jobs. The message of Larry P. is that psychologists must involve themselves in the task mandated in the last sentence of the court's opinion: "this will clear the way for more constructive educational reform" (quoted in Opton, 1979).

Others did not share the belief that the Larry P. decision was a social victory. Nadine Lambert, who was an expert witness for the state, felt it was a terrible decision. On learning of it, she remarked, "I think the people who will be most hurt by it are the Black children" (quoted in Opton, 1979, p. 1). Banning the use of IQ tests opens the door to completely subjective judgments, which may be even more racist than the test results. Opponents of the Larry P. decision cite many instances in which gifted black children were assumed to be average by their teachers but were recognized as highly intelligent because of IQ test scores.

The Larry P. decision has been frequently cited in subsequent cases. Some of these are actually remote from the issues in Larry P. For example, in the matter of Ana Maria R.*, parental rights were terminated on the grounds that the mother was mentally retarded. However, the mother was Spanish speaking and Larry P. was cited as precedent that tests and classification of mental retardation are discriminatory against blacks and Hispanics. In contrast to the case of Ana Maria R., the factual situation in an Illinois case was quite similar to Larry P. That case is described in the following section.

Parents in Action on Special Education v. Hannon.† Just as the case of Larry P. was making headlines in California, a similar case came to trial in Illinois. The case was a class-action lawsuit filed on behalf of two black children (representing the class of all similar children) who had been placed in special classes for the educable mentally handicapped (EMH) on the basis of IQ test scores. Attorneys for the two student plaintiffs argued that the children were inappropriately placed in EMH classes because of racial bias in the IQ tests. They suggested that the use of IQ tests for black children violates the equal protection clause of the Constitution and many federal statutes.

In their presentation to the court, the plaintiffs relied heavily on the recent Larry P. decision, which held that the WISC, the WISC-R, and the Stanford-Binet IQ tests are biased and inappropriate for the testing of minority children. However, Judge John Grady of the U.S. District Court came to exactly the opposite conclusion of Judge Peckham, who had presided over the Larry P. case. Judge Grady found evidence for racial bias in the three major IQ tests to be unconvincing. In his opinion, he noted that the items objected to were only a fraction of the items on the entire test. For example, witnesses for the plaintiffs never mentioned whole subtests on the WISC and WISC-R, such as arithmetic, digit span, block design, mazes, coding, and object assembly. The judge noted that these subtests were not biased in favor of either black or white children because most youngsters of both groups would have never confronted problems of this type before. The items for which there were legitimate objections were too few to have an impact on test scores.

Thus, less than one year after the historic Larry P. case, another court concluded, "Evidence of racial bias in standardized IQ tests is not sufficient to

*96 U.S. 2040(c)(1976).
†C.A. No. C-70 37 RFP (N.D. Cal., filed Feb 3, 1970).

render their use as part of classifications procedures to place black children in 'educable mentally handicapped' classes violative of statutes prohibiting discrimination in federally funded programs." In early 1984 the ninth district court of appeals considered an appeal citing both Larry P. and Parents in action. The court upheld Larry P. by a 2–1 vote (*Los Angeles Times,* January 24, 1984).

Debra P. V. Turlington.* Some people feel that a test is biased if it contains questions that particular test takers cannot answer. One 1979 lawsuit in Florida involved ten black students who had failed their first attempt to pass Florida's minimum competence test, the State Student Assessment Test. Debra P. was one of the students, and the case took her name. In Hillsborough County, where the suit was filed, about 19% of the students in the public school system were black. However, black students constituted 64% of those who had failed the test.

Minimum competence tests similar to the one used in Florida have been adopted by more than 30 states, and 19 states require the exam for graduation. If they meet other requirements, students who do not pass the exam are given a certificate of completion that acknowledges that they attended high school but does not carry the same status as a high school diploma. The Florida suit charged that the test should not be used for minority students when most of their education occurred before the schools were desegregated. Thus the dispute concerned whether the same test should be used for students who may have had unequal opportunities to learn in school. Attorneys for the students argued that their clients had been in inferior schools and had been the subjects of continued discrimination. Thus they should not be held to the standards for majority students, who had better opportunities.

Ralph D. Turlington was the commissioner of education and one of the defendants in the case. He argued that basic minimum standards must be applied in order to certify that students have enough information to survive in situations that require high school level sophistication. These standards, it was argued, must be absolute. Either students know the basic information or they do not. According to the commissioner, "To demand that a 12th-grade student with a 3rd-grade reading level be given a diploma is silly."

The Florida case illustrates the kind of lawsuits we might expect in the future. It pits two sides with reasonable arguments against each other. One side argues that minority children have worked hard in school under great disadvantage and cannot be expected to have learned the things majority children know. In recognition of their work they deserve a diploma. The other side argues that there should be an absolute standard for basic information (Seligmann, Coppola, Howard, & Lee, 1979).

The court essentially sided with the commissioner. The judge did not challenge the validity of the test. However, he did suspend the use of the test for four years, after which all the students who had any part of their education in

*347 F Supp. 1306 (N.D. Cal 1972). aff'd 502 F.2d 963 (9th Cir. 1979).

segregated schools would have graduated. Then, according to the opinion, the test could be used.

In a 1981 paper, Lerner argued that minimum competency exams, such as the SSAT II used in the State of Florida, benefit both students and society. As an attorney, she found little legal justification for court involvement. However, the court reopened the Debra P. case the same year as Lerner's paper was published. This new consideration came after those students who had begun their education under a segregated system had graduated and differences in performance could not be attributed to segregation. In the new evaluation, the U.S. district court of appeal considered the validity of the test. It stated that the test would violate the Equal Protection Clause if, "the test by dividing students into two categories, passers and failers, did so without a rational relation to the purpose for which it was designed, then the Court would be compelled to find the test unconstitutional" (474 F. Supp at 260). However, in this case, the Court concluded that the test did have adequate construct validity and that it could be used to evaluate functional literacy. In the same opinion, the Court stressed that the test must reflect what is taught in school and that continual surveillance of test fairness is warranted.

Regents of the University of California v. Bakke. Alan Bakke was an engineer in his thirties who decided to apply to the University of California, Davis, medical school in the early 1970s. Although Bakke had a high grade point average and good MCAT scores, he was denied admission. Bakke decided to investigate the matter. He discovered that his test scores were higher than those of minority students who had gained admission to the medical school under a special affirmative action program. Bakke eventually sued the university on grounds that he had been discriminated against because he was not a minority-group member. The suit ended in the Supreme Court and is considered to be one of the most important cases of the century.

Although many arguments were presented in the Bakke case, one of the major ones concerned the use of test scores. Under the affirmative action program, the cutoff value for MCAT scores was higher for nonminority than for minority students. In defense of the special admissions program it was argued that the tests were not meaningful (valid) for minority students. However, evidence was also presented that the tests were equally meaningful for both groups.

The Supreme Court ruling was not specific with regard to the use of tests. The court ruled that the university had to admit Bakke and that it had denied him due process in the original consideration of the case. It also implied that the use of different cutoff scores was not appropriate. However, the court did acknowledge that race could be taken into consideration in selection decisions. This acknowledgment was interpreted by the EEOC as meaning that affirmative action programs based on numerical quotas could continue (Norton, 1978).

After the Bakke decision, the high court seemed unwilling to hear further reverse-discrimination cases. For example, a week after Bakke, the court refused to hear a challenge to a strong affirmative action plan that created reverse discrimination (EEOC v. A.T. & T.).

Personnel Cases in Law. Most of the cases we have discussed involved educational tests. Several other important lawsuits have dealt with the use of tests in employment setting. Through a series of Supreme Court decisions, specific restrictions have been placed on the use of tests for the selection of employees. The most important of these cases are *Griggs v. Duke Power Company,** *Albemarle Paper Company v. Moody,*† and *Washington v. Davis.*‡ The effect of these decisions has been to force employers to define the relationship between test scores and job performance and to define the measure of job performance. However, none of the decisions denies that tests are valuable tools in the personnel field and that the use of tests can continue.

The courts have also been asked to decide on issues of test administration. For example, an employee of the Detroit Edison Company was not promoted because of a low test score. In his defense, his union suggested that the low score might have been an error and requested a copy of the test to check the scoring. Detroit Edison did not want to release the test because it feared that the union would distribute the items to other employees. By a vote of five to four, the Supreme Court ruled on the side of Detroit Edison (*Detroit Edison Co. v. NLRB*). It is interesting that in a major decision, such as this, a single vote can make a difference in policy (Cronbach, 1980).

A Critical Look at Lawsuits

The problems that psychologists are unable to resolve themselves will eventually be turned over to someone else for a binding opinion. This procedure may not be in the best interest of the field of psychology or of the people whom we serve.

It is difficult to be an uncritical admirer of the courts. As Lerner (1979) notes, inconsistencies in court decisions are commonplace. Even worse, judges who make important decisions about the use of tests often have little background in psychology or testing. Often judges obtain their entire education about testing during the course of a trial.

In the near future we must grapple with many tough issues. For example, many current social problems seem related to the differential distribution of resources among the races in American society. Changing the income distribution seems to be one of the only ways in which effective social change can occur. To accomplish this redistribution, we must get minority children in appropriate educational tracks, into professional schools, and into high-income positions. The courts have ruled that psychological tests are blocking this progress.

Psychologists themselves are not of one mind regarding the use of psychological tests. However, current research tends not to confirm the widely held belief that the tests are systematically biased. The field of psychometrics, after long and careful consideration of the problems, has come to a conclusion opposite

*414 NYS 2d. 982 (1979).
†74 C C3586, USDC (N.D. Ill, 7/7/80).
‡474 F. Supp. 244(M.D. Fla 1979) 644 F.2d 397 (1981).

to that of the courts, which have given the issue a briefer evaluation. In the end, however, the courts have the power, and their judgment is the law.

CONTROVERSIES SURROUNDING THE SOCIAL AND POLITICAL IMPLICATIONS OF TESTING

Psychological testing may lead to undesirable social and political trends. Some groups believe that psychological testing has produced social injustices that are indefensible in a free society. In this final section, two of these controversies will be discussed. The first is the review of the impact of psychological testing by Ralph Nader and his consumer organizations. The second stems from the accusation that psychological testing evidence was used as the basis for a racist immigration policy passed by the Congress in 1924.

Nader's Raid on the Educational Testing Service

Ralph Nader and his associates have established a sound reputation as consumer advocates. In 1980, Nader and junior associate Alan Nairn (1980) issued a report that criticized the testing industry. The prime target of the report was the Educational Testing Service (ETS). In a much publicized report, Nairn characterized ETS as an evil bureaucracy operating under a guise of secrecy. It was suggested that ETS conspired to maintain the status quo by intentionally discriminating against students from low-income families. It is suggested that invalid and biased tests ensure that only those from wealthy families are admitted to prestigious colleges and universities.

There are two major arguments in Nairn's report. First, the report suggests that the Scholastic Aptitude Test (SAT), which was developed by ETS, is not a valid predictor of college success. The second argument is that the only real correlate of the SAT is family income. Thus those selected for college admission through SAT testing programs may be wealthy but not necessarily more likely to succeed in college.

When carefully examined, many of the arguments in the Nairn report were found to be faulty. The report suggested that the SAT is no better than chance in predicting college performance. However, Nairn confused percentage of variance accounted for by the test with percentage of cases perfectly predicted. Reanalysis of the data revealed that the SAT predicts performance better than chance in nearly all cases. In combination with high school grades, the SAT is a relatively good predictor of college success for students from different social classes and with different ethnic group backgrounds. Although SAT scores are correlated with social class, there is no evidence for differential validity at different income levels (Kaplan, 1982).

The Nader report was very influential in the passage of some truth-in-testing legislation. As noted earlier, this legislation requires testing companies to make their items public. One consequence of this legislation is that test items may be

included in common instruments without having established a record of validity. The impact of the Nader proposal may serve to decrease the validity of the tests. In other words, the quality of the product may suffer. At the same time as the quality would decrease, costs are expected to increase due to the expense of continually creating and testing new items. Students might expect to pay a higher fee to take a less valid test. In other words, the Nader-Nairn report that was written in advocacy of the consumer may result in higher fees for a less valid product.

The Immigration Act of 1924

In 1924, the United States Congress passed a racially biased immigration law. This Immigration Act of 1924 set quotas for immigration based on the percentages of immigrants from each country who had arrived prior to the 1890 census. Most immigrants from Eastern and Southern Europe had arrived after 1890. As a result, the Immigration Act of 1924 had a strong bias in favor of immigrants from Northern and Western Europe and a bias against those from Southern and Eastern Europe.

Critics of intelligence tests have argued that testing advocates using data from biased intelligence tests played a central role in the passage of this act. This position was forcefully argued by Leon Kamin in a widely cited book entitled, *The Science and Politics of IQ* (Kamin, 1974, p. 16). Kamin deplored the "involvement of the mental testing movement in the passage of an overtly racist immigration act in 1924" (1982, p.16). A popular book by Paleobiologist Steven Gould (1981) entitled, *The Mismeasure of Man*, also attributes the passage of the Immigration Act of 1924 to testing advocates. Gould's book is widely quoted in both scientific and popular literature and has received overwhelmingly positive critical acclaim.

Recently, Snyderman and Herrnstein (1983) reexamined the role of intelligence test data in the passage of the 1924 Act. Evidence frequently cited by Kamin and his many followers include the report by H. H. Goddard (1917) characterizing as feebleminded the great majority of immigrants of Jewish, Hungarian, Italian, and Russian heritage. This attitude presumably led to the increased restrictions against immigration by Eastern and Southern Europeans. Kamin and Gould also cited C. Brigham's analysis of Army Alpha and Beta Intelligence Tests. Brigham (1923) concluded that the average test scores of immigrants from Northern Europe were better than those obtained from immigrants from Southern or Eastern Europe. In addition, an analysis of scores from draftees suggested that average intelligence scores had been declining for about a twenty-year period (in essence, from about the turn of the century). This decline could be accounted for by the increasing proportion of immigrants from Southern and Eastern Europe. Thus he recommended immigration policies that would favor Northern Europeans and restrict immigration from Southern and Eastern Europeans.

Brigham's position is clearly racist and difficult to defend. In addition, it appears that the Immigration Act of 1924 closely coincides with Brigham's

recommendations. Nevertheless, Snyderman and Herrnstein's (1983) careful evaluation suggests that Brigham's statements had relatively little influence in the deliberations of the Congress. Critics of intelligence tests, including Kamin and Gould, imply that Brigham's analysis went essentially unchallenged. Yet there is substantial evidence that psychologists of the day were highly critical of Brigham's work and of his position on immigration. In fact, Brigham himself changed his position a few years later (Brigham, 1930).

In a similar vein, it appears that Goddard's (1917) views were largely misrepresented. Goddard had intentionally preselected his sample to include only those of borderline intelligence. In his study of 178 immigrants, there was no intention to make statements about the prevalence of feeble-mindedness among immigrants in any particular group. The purpose was to demonstrate that a test could make fine discriminations in borderline cases. Yet it was well known at the time that the test exaggerated the rate of feeble-mindedness in all adult populations. There is also evidence, contrary to statements by Kamin (1974) and Gould (1981), that Goddard did not attribute poor intelligence test performance to either inheritance or ethnic background. In fact, Goddard attributed poor performance among some immigrants to poor environment (Snyderman and Herrnstein, 1983). The views of both Goddard and Brigham were widely criticized by the time the immigration policies of 1924 were formulated. Gould's book (1981) stated that intelligence test data and the consensus among psychologists were the focal point of Congressional deliberation. As Gould (1981) noted, "Congressional debates leading to the passage of the Immigration Restriction Act of 1924 continually invokes the Army data (p. 232). Upon review of Congressional Record, Snyderman and Herrnstein were unable to substantiate these claims. In fact, the 32 sections of the act make no reference to intelligence tests, intelligence, or feeblemindedness. Kamin (1974) suggested that testimony and written documents had influenced the members of the Congressional committees prior to the floor debates. For example, Kamin quoted a statement from Madison Grant (1916), an anthropologist who praised the value of the Army Intelligence tests. Yet Snyderman and Herrnstein were unable to locate Grant's statement in the Harvard University archives and could find no record that the Senate Immigration Committee had even met on the day Grant proportedly made his statement. In fact, there was no evidence that Grant had ever made a statement to the Committee.

The major advocates of psychological tests, including Goddard, Termin, Yerkes, and Thorndike were never even called to testify. The records show that those few witnesses who mention native differences in intellect were typically criticized by the members of Congress when they presented their testimony. In summary of their review, Schnyderman and Herrnstein (1983) stated, "Summarizing our examination of the Congressional Record and Committee hearings: there is no mention of intelligence testing in the act; tests results of the immigrants appear only briefly in the committee hearings and are largely ignored or criticized, and they are brought up only once in the over 600 pages of the Congressional floor debate where they are subjected to further criticism without rejoinder" (p.

994). Thus there is little evidence supporting the commonly held belief that the unified group of psychologists used intelligence test data to influence the regrettable Immigration Act of 1924.

SUMMARY

This chapter has reviewed several controversies in testing. Most of these controversies are related to differential performance between ethnic and racial groups on standardized tests. At present there is little convincing evidence that content bias in major tests is the major cause of observed differences. Studies of criterion validity are more difficult to evaluate. Interpretation of the results depends on a specific philosophical orientation. Thus disagreements may be a reflection of different moral positions.

Several alternative approaches have been proposed. At present each is still under evaluation. While these issues are debated in academic circles, a greater number of cases has reached the courts. Future court battles can be anticipated, since there has been considerable inconsistency in judgments. Social and political debates about testing have often been emotional and have gained considerable public attention. However, some of the claims have not been well substantiated and many of the anti-testing public statements appear to have been in error.

REFERENCES

Anastasi, A. (1980). Abilities and the measurement of achievement. *New Directions for Testing and Measurement, 5,* 1–10.

Andre, J. (1976). Bi-Cultural Socialization and the Measurement of Intelligence. *Dissertation Abstracts International,* 3676B–3676B.

APA (1983). American psychological association statement on test item disclosure legislation (mimeograph).

Bartlett, C. J., & O'Leary, B. S. (1969). A differential prediction model to moderate the effects of heterogeneous groups in personnel selection and classification. *Personnel Psychology, 22,* 1–17.

Bergan, J. R., & Parra, E. B. (1979). Variations in IQ testing and instruction and the letter learning and achievement of Anglo and bilingual Mexican-American children. *Journal of Educational Psychology, 71,* 819–826.

Bersoff, D. N. (1979). Regarding psychologists testily: Legal regulation of psychological assessment in the public schools. In B. Sales & M. Novick (Eds.), *Perspectives in law and psychology. III: Testing and evaluation.* New York: Plenum.

Bersoff, D. N. (1981). Testing and the law. *American Psychologist, 36,* 1047–1057.

Bianchini, J. C. (1976, May). *Achievement tests and differentiated norms.* Paper presented at the U.S. Office of Education invitational conference on achievement testing of disadvantaged and minority students for educational program evaluation, Reston, Va.

Brigham, C. C. (1923). *A study of American intelligence.* Princeton, NJ: Princeton University Press.

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review, 37,* 158–165.

Brill, S. (1973). The secrecy behind the college boards. *New York Magazine.* (Reprinted by the NYC Corporation.)

Brown, F. G. (1979a). The algebra works—But what does it mean? *School Psychology Digest, 8*(2), 213–218.

Brown, F. G. (1979b). The SOMPA: A system of measuring potential abilities? *School Psychology Digest, 8*(1), 37–46.

Clarizio, H. F. (1979a). In defense of the IQ test. *School Psychology Digest, 8*(1), 79–88.

Clarizio, H. F. (1979b). SOMPA—A symposium continued: Commentaries. *School Psychology Digest, 8*(2), 207–209.

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

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

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

Cole, N. S. Bias in testing. (1981). *American Psychologist, 36,* 1067–1077.

Copple, C. E., & Succi, G. J. (1974). The comparative ease of processing standard English and Black nonstandard English by lower-class Black children. *Child Development, 45,* 1048–1053.

Cordes, C. (1983). Jensen Refines Theory Linking G-Factor to Processing Speed. *APA Monitor, 14*(10), 3,20.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30,* 1–14.

Cronbach, L. J. (1978). Black Intelligence Test of Cultural Homogeneity: A review. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (vol. 1). Highland Park, NJ: Gryphon Press, 1978.

Cronbach, L. J. (1980). Validity of parole: How can we go straight? *New Directions for Testing and Measurement, 5,* 99–108.

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

Darlington, R. B. (1978). Cultural test bias: Comment on Hunter and Schmidt. *Psychological Bulletin, 85,* 673–674.

Das, J. P., Kirby, J., & Jarman, R. F. (1975). Simultaneous and successive syntheses An alternative model for cognitive abilities. *Psychological Bulletin, 82,* 87–103.

Das, J. P., Kirby, J., & Jarman, R. F. (1979). *Simultaneous and successive cognitive processes.* New York: Academic Press, 1979.

Dobko, P., Kehoe, J. F. (1983). On the Fair Use of Bias: A Comment of Drasgow. *Psychological Bulletin, 93,* 604–408.

Drasgow, F. Biased Test Items and Differential Validity. (1982). *Psychological Bulletin, 92,* 526–531.

EEOC (1970). Equal Employment Opportunity Commission guidelines on employee selection procedures. *Federal Register, 35* (19), 12333–12336.

EEOC (1978). *Employment Guidelines,* Washington, D.C.: U.S. Government Printing Office.

Flaugher, R. L. (1974). *Bias in testing: A review and discussion* (TM Rep. 36). Princeton, NJ: ERIC Clearinghouse on Tests, Measurements, and Evaluation.

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

Flaugher, R. L., & Schrader, W. B. (1978). *Eliminating differentially difficult items as an approach to test bias* (RB-78-4). Princeton, NJ: Educational Testing Service.

Garcia, J. (1981). The logic and limits of mental aptitude testing. *American Psychologist, 36,* 1172–1180.

Goddard, H. H. (1917). Mental tests and the immigrant. *Journal of Delinquency, 2,* 243–277.

Goodman, J. (1977). The diagnostic fallacy: A critique of Jane Mercer's concept of mental retardation. *Journal of School Psychology, 15,* 197–206.

Goodman, J. (1979). "Ignorance" versus "stupidity"—the basic disagreement. *School Psychology Digest,* 1979, 8(2), 218–223.

Gordon, E. W., & Terrell, M. D. (1981). The changed social context of testing. *American Psychologist, 36,* 1167–1171.

Gould, S. J. (1981). *The Mismeasure of Man,* New York: Norton.

Grant, M. (1916). *The passing of the great race.* New York: Scribner's, 1916.

Green, D. R., & Draper, J. F. (1972, September). *Exploratory studies of bias and achievement Association,* Honolulu, Hawaii.

Guion, R. M. (1976). Recruiting, selection and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally.

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

Haney, W. (1981). Validity, vaudeville, and values: A short history of social concerns over standardized testing. *American Psychologist, 36,* 1021–1034.

Hardy, J. B., Welcher, D. W., Mellitis, E. D., & Kagan, J. (1976). Pitfalls in the measurement of intelligent: Are standarized intelligence tests valid for measuring the intellectual potential of urban children? *Journal of Psychology, 94,* 43–51.

Hodos, W., & Campbell, B. C. G. (1969). Scala naturae: Why there is no theory in comparative psychology. *Psychological Review, 76,* 337–350.

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of statistical and ethical implications of various definitions of test bias. *Psychological Bulletin, 83,* 1053–1071.

Hunter, J. E., & Schmidt, F. L. (1978). Bias in defining test bias: Reply to Darlington. *Psychological Bulletin, 85,* 675–676.

Ironson, G. H., & Sebkovial, N. J. (1979). A Comparison of Several Methods for Assessing Item Bias. *Journal of Educational Measurement, 16,* 209–225.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1–23.

Jensen, A. R. (1972). *Genetics and education.* New York: Harper & Row.

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

Jensen, A. R. (1983, August). *The nature of black–white difference on various psychometric tests.* Presented at the Meeting of the American Psychological Association Anaheim, California.

Jones, L. V. (1983). White–black achievement differences: The narrowing gap. Invited Address, American Psychological Association, Anaheim, CA, August.

Kagan, J., Moss, H. A., & Siegel, I. E. (1963). Psychological significance of styles of conceptualization. *Monographs of the Society for Research in Child Development, 28*(2, Serial No. 86), 73–124.

Kallingal, A. (1971). The prediction of grades for black and white students at Michigan State University. *Journal of Educational Measurement, 8,* 263–265.

Kamin, L. J. (1974). *The science and politics of IQ.* Hillsdale, NJ: Erlbaum.

Kamin, L. J. (1976). Heredity, intelligence, politics and psychology: II. In N. J. Block & G. Dworkin (Eds.), *The I.Q. controversy.* New York: Pantheon Books.

Kamin, L. J. (1982). Mental testing and immigration. *American Psychologist, 37,* 97–98.

Kaplan, R. M. (1982). Nader's Raid on the testing industry: Is it in the best interest of the consumer? *American Psychologist, 37,* 15–23.

Kaplan, R. M. & Saccuzzo, D. P. (1982). *Psychological testing: Principles, Applications and Issues.* Monterey: Brooks/Cole.

Kaufman, A. S., & Kaufman, N. L. (1983). *K-ABC Kaufman Assessment Battery for Children.* Circle Pines, Minnesota: American Guidance Service.

Kaufman, A. S., & Doppelt, J. D. (1976). Analysis of WISC-R standardization data in terms of the stratification variables. *Child Development, 47,* 165–171.

Kiersh, E. (1979, January 15). Testing is the name, power is the game. *The Village Voice.*

Lerner, B. (1979). Tests and standards today: Attacks, counterattacks, and responses. *New Directions in Testing and Measurement, 1*(3), 15–31.

Lerner, B. (1981). The minimum competence testing movement: Social, scientific, and legal implications. *American Psychologist, 36,* 1057–1066.

Lesser, G. S., Fifer, G., & Clark, D. H. (1965). Mental abilities of children from different social-class and cultural groups. *Monographs of the Society for Research in Child Development, 30*(4, Serial No. 102).

Levy, S. (1979). E.T.S. and the "coaching" cover-up. *New Jersey Monthly, 3*(5), 4–7.

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. (1975). *Race Differences and Intelligence.* San Francisco: Freeman.

Long, P. A., & Anthony, J. J. (1974). The measurement of retardation by a culture-specific test. *Psychology in the Schools, 11,* 310–312.

Luria, A. R. (1966). *Human brain and psychological processes.* New York: Harper & Row.

Luria, A. R. (1966). *Higher Cordical Functions in Man.* New York: Basic Books.

Mannheim, K. (1936). *Ideology and utopia.* London: Kegan, Paul, Trench, Trubner.

McCormick, E. J., & Ilgen, D. (1980). *Industrial psychology* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Mercer, J. R. (1971). Sociocultural factors in labeling mental retardates. *Peabody Journal of Education, 48,* 188–203.

Mercer, J. R. (1973). *Labeling the mentally retarded: Clinical and social system perspective on mental retardation.* Berkeley: University of California Press.

Mercer, J. R. (1972). Anticipated achievement: Computerizing the self-fulfilling prophecy. Presented at the meeting of American Psychological Association, Honolulu.

Mercer, J. R. (1979). In defense of racially and culturally non-discriminatory assessment. *School Psychology Digest, 8*(1), 89–115.

Mercer, J. R., & Lewis, J. F. (1979). *System of multi-cultural pluralistic assessment: Conceptual and technical manual.* New York: Psychological Corporation.

Milgram, M. A. (1974). Danger: Chauvinism, scapegoatism, and euphenism. In G. J. Williams & S. Gordon (Eds) *Clinical Child Psychology: Current Practices and Future Perspectives,* New York: Behavioral Publications.

Munsinger, H. (1975). The adopted child's IQ: A critical review. *Psychological Bulletin, 82,* 623–659.

Nairn, A., & Associates. (1980). *The reign of ETS: The corporation that makes up minds.* Washington, DC: Nader.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Norton, E. H. (1978, July). *The Bakke decision and the future of affirmative action.* Statement of the Chair, U.S. Equal Employment Opportunity Commission, at the National Association for the Advancement of Colored People convention.

Oakland, T. (1979). Research on the ABIC and ELP: A revisit to an old topic. *School Psychology Digest, 8,* 209–213.

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

Opton, E. A. (1977). A psychologist takes a closer look at the recent landmark Larry P. opinion. *APA Monitor,* December, 1–4.

Opton, E. (1979, December). A psychologist takes a closer look at the recent landmark Larry P. opinion. *APA Monitor,* pp. 1,4.

Ornstein, R. (1972). *The psychology of consciousness.* San Francisco: Freeman.

Pettigrew, T. F. (1964). *A profile of the American Negro.* New York: Van Nostrand Reinhold.

Pfeifer, C., & Sedlacek, W. (1971). The validity of academic predictor for black and white students at a predominantly white university. *Journal of Educational Measurement, 8,* 253–261.

Piersel, W. C., Blake, B. S., Reynolds, C. R., and Harding, R. (1982). Bias and content validity of the Boehm Test of basic concepts for white and Mexican-American children. *Contemporary Educational Psychology, 7,* 181–189.

Quay, L. C. (1971). Language dialect, reinforcement, and the intelligence-test performance of Negro children. *Child Development, 42,* 5–15.

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

Reschly, D. J. (1981). Psychological testing in educational classification and placement. *American Psychologist, 36,* 1094–1102.

Reynolds, C. R. (1980). An examination of bias in the pre-school test battery across race and sex. *Journal of Educational Measurement, 17,* 137–146.

Reynolds, C. R., & Nigl, A. J. (1981). A regression analysis of differential validity: An intellectual assessment for black and white inner-city children. *Journal of Clinical and Child Psychology, 10,* 176–179.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom.* New York: Hold, Rinehart & Winston.

Sattler, J. M. (1979, April). *Intelligence tests on trial; Larry P. et al. vs. Wilson Riles et al.* Paper presented at the meeting of the Western Psychological Association, San Diego.

Sattler, J. M. (1982). *Assessment of children's intelligence and special abilities.* Boston: Allyn & Bacon.

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

Scheuneman, J. D. (1981). A new look at Bar S and aptitude tests. *New Directions in Testing and Measurement,* No. 12, 5–25.

Seligmann, J., (1979, May 28). Coppola, V., Howard, L., & Lee, E. D. (1979, May 28). A really final exam. *Newsweek,* pp. 97–98.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190.

Snyderman, M., & Herrnstein, R. J. (1983). Intelligence tests and the Immigration Act of 1924. *American Psychologist, 38,* 986–995.

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

Taking the Chitling Test. (1968, July 15). *Newsweek,* pp. 51–52, 72.

Temp, G. (1971). Test bias: Validity of the SAT for blacks and whites in thirteen integrated institutions. *Journal of Educational Measurement, 8,* 245–251.

Thorndike, R. L. (1968). Review of *Pygmalion in the Classroom* by R. Rosenthal and L. Jacobson. *American Educational Research Journal, 5,* 708–711.

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

Williams, R. L. (1972, September). *The BITCH-100: A culture-specific test.* Paper presented at the Meeting of the American Psychological Association, Honolulu, Hawaii.

Williams, R. L. (1974). Scientific racism and IQ: The silent mugging of the black community. *Psychology Today, 7,* 32–41.

Woodring, P. (1966). Are intelligence tests unfair? *Saturday Review, 49,* 79–80.

Zores, L. S., & Williams, P. B. (1980). A look at content bias in IQ tests. *Journal of Educational Measurement, 17,* 313–322.