

Test Bias or Cultural Bias: Have We Really Learned Anything?

by

Stephen J. Schellenberg  
Saint Paul (MN) Public Schools

A paper presented as part of the symposium  
“The Achievement Gap: Test Bias or School Structures?”  
sponsored by the National Association of Test Directors as part of the  
Annual Meeting of the National Council for Measurement in Education  
San Diego, California, April 14, 2004

Running Head: Test bias or cultural bias

### Abstract

Standardized achievement testing regularly shows large gaps between the scores of different groups – Caucasian vs. students of color, middle class vs. low income, etc. Those who would study this gap are confronted with a paradox. Are the gaps the result of true differences in achievement or are they the result of bias in the measurement instruments themselves? Each of these explanations appears to preclude the other. This paper examines psychometric and socio-cultural approaches to this question in an attempt to identify several overarching issues that need to be addressed by test directors to optimize the value of achievement testing for students. It also examines recent developments, including some implications of the No Child Left Behind law.

## Introduction

This year, we celebrate the fiftieth anniversary of the Supreme Court decision in the case of *Brown vs. Board of Education*. Like all great Supreme Court decisions, it brought about fundamental changes in the country that were felt well beyond the specific focus of the case – equal access to public education. The case was not argued using achievement test data to show unequal results, as it might be today. Indeed, when the case was being argued in 1952 and 1953, standardized achievement testing was in its infancy. Machine-readable scan sheets were still an experiment and nationally normed achievement tests were largely unknown. Rather, the case was argued from a standpoint of equal treatment, that separate could never be equal. If the fallout from *Brown* had been limited to equal access, we would not be discussing achievement gaps and test bias today. However, the expectation of equality of access led to an equality of hope and a reasonable expectation that there would ultimately be equality of results. Rightly or wrongly, the documentation of these results is one of the major roles adopted by standardized achievement tests.

As standardized achievement tests became a fact of American educational practice, they showed massive gaps between racial groups, social strata, and regions of the country. At first, these were taken to be evidence of unequal educational systems, but, as educational access became more equitable and the score gaps remained, researchers and activists began to look for other explanations. Could it be that the tests themselves were faulty? If the tests did not fairly measure the accomplishments of some groups, what good were they? Educators who were intent on improving the lot of historically underserved classes were left in the paradoxical position of arguing that test data showed that certain groups of students

were not achieving, but that the test was itself unfair and therefore not to be trusted. This paradox is still alive today and is, in fact, the subject of this symposium.

Scholarly research on test bias has developed along two broad lines – psychometric and socio-cultural, receiving its highest exposure during the 1970's and early 1980's (cf. Berk, 1982; Jensen, 1980; Reynolds and Brown, 1984). Test publishers in turn incorporated controls for bias into their test development procedures as a matter of course, so that the most obvious biases are no longer evident. To a large degree, the subject of test bias has become the province of policy debates among politicians and social activists and increasingly technical discussions among psychometricians. Although it no longer occupies center stage as a research topic, those of us who work in testing should not be lulled into a false sense of calm. The issues raised in the earlier go-around have not been fully addressed and testing looms so large in the public policy arena that we are sure to see a reemergence of the debate regarding fairness in educational testing.

In that policy arena, the conversation is becoming increasingly dominated by polemic rather than systematic inquiry. The polemical positions emerge from hardened points of view that see no middle ground. At one extreme are those who insist that group differences are *prima facie* evidence that the tests and everything else in public education are biased; at the other extreme are those who insist that different test scores tell the truth and anyone who seeks a more nuanced explanation is simply making excuses for poor performance. It is absolutely essential that the conversation be informed by thoughtful research. In the words of Lee Cronbach,

Public controversy deals with stereotypes, never in subtleties. The Luddites can smash a device but to improve a system requires calm study . . . . Sound policy is not for tests or

against tests, but how tests are used. But all the public hears is endless angry clamor from extremists who see no good in any test, or no evil.” (1975)

This paper will trace the evolution of some of the many perspectives on test bias and suggest some ways in which they can be reconciled and combined. Several different approaches to the subject will be outlined, both within psychometric and socio-cultural viewpoints. We will examine some recent developments in large-scale assessment and speculate on the effects of high-stakes testing that is mandated by the No Child Left Behind Act. We will primarily discuss standardized achievement testing in schools, rather than other forms of testing such as psychological assessments or workplace testing. Most of all, we are seeking a practical, yet defensible stance for testing professionals in the field as they deal with this complex issue.

### Different approaches to differential performance on tests

#### Early history

From the beginning of modern psychological testing, researchers have found differential results. In the work of Binet and the early development of the Army Alpha, group differences were noted and even assumed (Matarazzo, 1972). Eels et al (1951) summarize the three general explanations that were common at the time:

1. Subjects scoring well on tests are genuinely superior in inherited genetic equipment.
2. Large-scale group differences are probably the result of a faulty test.
3. High scores are generally the product of a superior environment and low scores the product of a poor one.

To this day, most investigations of differential achievement fall into one or the other of the above general statements.

We will not discuss the first of these explanations at any length. This is not to imply that it is not worthy of investigation. Rather, we are evading that discussion because it would lead down a very emotionally loaded and non-productive sidetrack (witness the reactions evoked by Jensen, 1980 and Herrnstein and Murray, 1994). As educators, we are charged with teaching all students, so to assume that their achievement is predestined through genetics is to assume that our job can't be done before we even start. The last two of the above explanations, however, are capsule summaries of the arguments that have led to most of the research in this field. The research itself falls into two general categories – psychometric and socio-cultural. Psychometric approaches concentrate on examining the testing instrument and students' responses to it. Socio-cultural approaches look at performance on the test as part of the overall context in which a student lives and learns. Rather than being incompatible viewpoints, as they are sometimes portrayed, these two approaches are complementary. Neither offers a complete picture, but both offer pieces of the total mosaic.

#### The psychometric framework.

Major challenges to test fairness emerged in the late 60's and early 70's as the result of the convergence of several factors. First, there was the emerging expectation of equality of results that we have already mentioned. In a parallel track was the developing view of African-American culture as an equal culture to the dominant white culture, and the resultant validation of approaches that emphasized its distinctive qualities. A very important factor

within this development was the emergence of a black psychology that attempted to study and assess African Americans within their own culture (Hilliard, 1995). Key tenets of this psychology were that assessment must be expressed through cultural material familiar to the test taker, and that assessment must take into account histories of oppression, including psychological oppression. Certainly, such a psychology could not rely on instruments developed under the old paradigm. Given the purposes of this movement, much of its attention on testing was focussed on psychological assessments and the cultural biases of such tests as IQ tests or personality inventories. However, this attention to testing inevitably spilled over into achievement testing as well.

In psychometric terms, test bias is fundamentally a validity issue. If a test cannot be trusted with identifiable subpopulations, it is not valid in that context. Four aspects of validity seem to have attracted the most attention:

1. content validity
2. construct validity
3. predictive validity
4. consequential validity

Within any of these aspects of validity, we must remember that validity exists only within a specific purpose for the test. Thus, a test may stand up well to validity challenges in one context and not in another.

The most studied of these four is content validity, where research has focussed nearly exclusively on item bias. Every national testing company can tell you in considerable detail the steps, both subjective and statistical, that have been taken to seek out and destroy biased

items. Subjective techniques usually involve panels of experts from diverse backgrounds examining items to detect potential bias. Most, if not all, of the statistical approaches share a common conceptual base. Higher or lower scores by a group on a given item are not sufficient evidence to identify a biased item. It may be that the underlying ability is actually different among the groups. Therefore, the group's performance on the item must be either better or worse than the group's performance on the test as a whole for the item to be eliminated. This extensive focus on item characteristics has led to fundamental changes in the appearance of standardized achievement tests. These changes have ranged from the obvious – inclusion of persons from multiple cultures in the illustrations and stories in the test – to much more subtle changes, such as elimination of references to products that may be common in one part of the country but not another.

The focus on item bias may have obscured our view of threats to construct validity, where psychometric tools alone cannot answer all the criticisms. Certainly, if a test measures different things in different populations, it lacks construct validity. A simple example would be a mathematics test that includes many complex word problems. The reading load may be so heavy that it actually measures reading comprehension for second language speakers. If a factor analysis finds different factor structures for different populations, we suspect that the test is not measuring the same construct across populations. However, statistical analysis rarely answers the critics who find fault with construct validity. Their criticisms often question the validity of the constructs themselves, not just whether the construct is being assessed equivalently. These questions about the purposes and values of public education are more appropriately the province of curriculum developers, policy makers and philosophers than of psychometricians.

Predictive validity, the accuracy of a test in predicting an outcome variable, has also generated its share of controversy. From a standpoint of predictive validity, the selection of the outcome measure is paramount. In general, achievement tests are accurate predictors of future academic success, but this fact may lead us into a circular argument. If we assume the outcome measure to be a truly independent measure, a positive correlation between the test and the criterion establishes predictive validity. On the other hand, if we assume the outcome measure to be merely a different form of the predictor, we have demonstrated nothing at all. All we have found is that one test is an accurate predictor of another test. This line of argument is frequently raised by social activists to illustrate their view that the entire educational system is biased in favor of middle-class, largely Eurocentric views. It is difficult to conceive of a psychometric approach that would either prove or disprove this argument. Once again, we find ourselves in territory that is much more comfortable for policy makers and philosophers than for testing specialists.

Consequential validity is closely related to predictive validity, but has more to do with the decisions that arise from test results. This form of validity is of particular concern in psychological and workplace testing, where the consequences for the test-taker can literally be life-changing. A frequent argument regarding the consequential validity of achievement tests occurs when tests are used to diagnose academic weakness and direct corrective action, an approach that test directors usually encourage. If those diagnoses and actions also lower our expectations for those students, they contribute to further academic weakness and may actually hold the student back. In the most dramatic case, a student may be directed toward or away from certain academic paths in ways that can be as life-changing as an employment test. To date, most discussions of consequential validity have focussed on the consequences

of tests for individuals, but this may change with the serious consequences for schools mandated by the No Child Left Behind Act.

Three large issues emerge from psychometric research into bias as a validity issue. First, the extensive focus on a rather narrow definition of content validity, i.e. item bias, may have prevented us from seeing threats to other forms of validity. Second, there is an uneasy balance between a test being indicative of individual performance versus group performance. If a test is believed to be biased against a group, is every member of that group considered an exemplar of the bias and therefore incapable of being assessed accurately? Finally, despite all the attention given to item bias in standardized tests, achievement gaps have narrowed very little and are diminished only slightly even if we hold family income and parental education constant (Jencks & Phillips, 1998). Clearly, psychometric analyses have not yet yielded the answers to these issues.

### The socio-cultural framework

An alternate view to the psychometric approach places testing as part of the cultural phenomenon of public schooling, which itself is reflective of larger societal and cultural issues. Frisby (1998) articulates three possible professional approaches to culture:

1. the theorist-researcher,
2. the practitioner-clinician, and
3. the socially conscious advocate.

It is within this framework that we may better be able to understand and respond to issues of cultural bias in testing. Psychometric analysis does not situate our work in a cultural context.

Without a cultural context, we cannot truly address cultural bias. Psychometric analyses may detect the artifacts of bias, but do little to explain or alleviate it.

The theorist-researcher seeks to understand the determinants of performance through objective development and evaluation of theory. In the present context, theorist-researchers probe the relationships among subgroup membership and academic performance, guided primarily by empirical data. However, because they are examining culture, they cannot do this by looking at test data alone. The question is framed within a broader view of culture, particularly bias toward a dominant culture. This study of bias begins with the curriculum, then looks to the ways in which knowledge and skills are assessed. Test data help define the issue, but the goal is to develop a better theory of what and how students learn and how this is affected by cultural background. Closely intertwined with this analysis is the question, “What must we ensure that all students learn?” In order to know what students have learned, we need to be sure that we are using assessment tools are congruent with both the cultural background and the desired outcomes.

The second professional role that Frisby identifies, the practitioner-clinician, comes closest to the test director’s role. The practitioner-clinician seeks reliable knowledge to guide practice. The approach is pragmatic, informed by theory, but oriented to practical solutions to everyday problems. This role draws on both psychometric and socio-cultural frameworks to make testing a tool toward student learning rather than an end in itself. In addition to seeking reliable knowledge to guide practice, the good practitioner-clinician will also see that test results are interpreted within a context that includes the culture of the student and the cultural assumptions under which the test was developed. This interpretation will be guided by beliefs about the set of skills that a student needs to acquire, regardless of cultural heritage.

The third role, the socially conscious advocate, seeks to protect rights, particularly of historically-excluded groups. As with the theorist-researcher, the advocate's position begins by examining the larger culture and the student's place within it, then moves to what students need to know and how we should assess it. However, the advocate is not concerned so much with objectivity or theory development as with correcting historic wrongs of oppression and exclusion. Because of this focus, practitioners may tend to dismiss advocates for not being well grounded in theory, but we owe much to their insistent voice. They have often forced the issues of achievement gaps and bias into the forefront when others, taking a seemingly more objective view, were willing to explain them away. Theirs is the balancing voice to the purely psychometric or theoretical view.

### Recent developments

The context of the test bias discussion has changed considerably since the 60's and early 70's. At that time, the issue was almost completely entwined with desegregation concerns. Several other issues have broadened the discussion in recent years. Court decisions regarding psychological testing for Special Education are having an influence on the uses of achievement tests as well, dictating a great deal more caution in applying test results beyond their intended uses. Adding to the issue of racial bias is concern about fairness of tests for second language speakers and students with handicaps. Test developers have responded to these concerns with untimed tests and sometimes with translated tests, each of which present new psychometric issues in their standardizations.

The format of achievement tests is gradually changing in other ways as well, moving away from nearly exclusive reliance on multiple choice formats into a variety of free-

response forms of assessment. In the mid-90's, performance assessment or "authentic" assessment was widely touted as a more complete way of measuring student achievement. However, in terms of eliminating test bias, it has so far shown rather disappointing results. In fact, in some cases, performance assessments show even wider achievement gaps than do multiple-choice formats. This may be because performance assessment relies heavily on expert judgment for its results and human judgment is notoriously difficult to standardize. On the other hand, multiple-choice formats may have understated the true extent of the achievement gaps, which are now revealed by the new assessments. From a practitioner's standpoint, performance assessment is very time-consuming and expensive to implement on a large scale. It has not yet shown its value as a tool to eliminate test bias, but has definitely expanded the practitioner's toolkit

One of the most far-reaching recent developments in testing is the implementation of the No Child Left Behind Act (NCLB). Several provisions of that law have very strong implications for discussions of test bias. The law asserts that all targeted subgroups (five racial/ethnic groups, Limited English Proficient students, students with handicaps, and low income students) must achieve at high standards. However, this requirement can only be supported by test results if those tests treat all subgroups fairly. The law also mandates that each state develop its own standards and its own tests. It remains to be seen whether these state-developed tests will take adequate precautions against test bias. Many states are simply contracting out their test development to established national firms with well-developed bias-reduction techniques. If states choose to develop their own tests in house, this aspect of test development may not be adequately addressed.

## Discussion

Achievement gaps remain a fact of American education. They are not simply an artifact of testing, as they appear in multiple other forms – graduation and dropout rates, participation in remedial programs, and rates of college completion, to name a few. Multiple indicators show that identifiable groups of students are not achieving as well as others. This is clearly not simply a problem of mismeasurement of students. It is a systemic problem in American education with deep roots in the society and multiple cultures in which our students live. The solution will not come from accusations that this test or that curriculum have inherent bias toward a dominant culture but through careful, thoughtful collaboration among concerned researchers, practitioners and policy-makers from multiple fields. All perspectives, psychometric and socio-cultural, need to be part of this discussion. The two central questions that must be addressed are the following:

1. What must we as educators ensure that students learn?
2. How far are we capable of stretching to accommodate diverse populations before we no longer teach students what they need?

The answers to those two questions must guide what and how we test.

Psychometric approaches to eliminating test bias have been relatively restrictive to date, looking primarily at item bias. While this approach has definitely changed the look and feel of standardized tests, it is important that we look beyond item bias to find some way of applying rigorous tests for more broad-ranging types of bias. On the practical side, we must continue our efforts to ensure that tests are used appropriately in contexts where they have validity. We must not resort to the glib mantra that “It’s not the test, but the use of the results, that is biased,” but must rather always seek an appropriate balance between what students

bring with them and what we as educators need to assure that they take away. We must also be willing to step forward when psychometric arguments are applied in inappropriate contexts.

Two equal and opposite dangers are presented in the question of racial and cultural bias in testing. The first is that we will believe that the problem has been solved, that bias in achievement testing has been minimized, and that nearly all remaining differences in test scores are real. This would imply that the vision of *Brown vs. Board of Ed* has been realized – that there is true equality of access and equality of hope for all students. The second, opposite danger is that we will explain away real differences as being the result of race or culture. This can be equally damaging to students, for it could lead to individual needs being undiagnosed and opportunities denied. Both positions deny the complexity of the role of culture in our society and both embody the bias that they try to reject.

Finally, there is considerable risk that the ongoing efforts to eliminate bias in testing will be derailed by the NCLB juggernaut. This risk takes two forms. First, the level of testing mandated in the Act could keep testing professionals so busy that the question of test bias simply slides off the plate. Second, there is the danger that extreme policy positions will dominate the discussion. One extreme assumes that standards define content that all students must learn, and that tests measure that learning accurately. The other extreme challenges both of those assumptions and asserts that schools are so impossibly biased and give such biased tests that we can't possibly expect equal results. Somewhere in the middle must sit the test director, trying to make tests work for students, not against them, but recognizing the psychometric and cultural forces inherent in any test.

References

- Berk, Ronald A. (1982), ed. *Handbook of methods for determining test bias*. Baltimore: Johns Hopkins University Press.
- Brown v. Board of Education, 347 US 483 (U. S. Supreme Court, 1954). Argued December 9, 1952. Reargued December 8, 1953.
- Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist*, 30, 1-14.
- Eels, K., Davis, A. Havighurst, R. J. Herrick, V. E., and Tyler, R. W. (1951) *Intellectual and cultural differences: a study of cultural learning and problem-solving*. Chicago: University of Chicago Press.
- Frisby, Craig L. (1998) Culture and cultural differences. In Sandoval, Jonathan et al, eds. *Test interpretation and diversity: achieving equity in assessment*. Washington, DC: American Psychological Association.
- Herrnstein, Richard and Murray, Charles (1994). *The bell curve: intelligence and class structure in American life*. New York: Free Press.
- Hilliard, Asa G., III (1995), ed. *Testing African American students: special reissue of the Negro Educational Review*. Chicago: Third World Press.
- Jencks, Christopher and Phillips, Meredith (1998), eds. *The black-white test score gap*. Washington, DC: Brookings Institution Press.
- Jensen, Arthur R. *Bias in mental testing*. New York: Free Press, 1980.
- Matarazzo, Joseph. (1972) *Wechsler's measurement and appraisal of adult intelligence*. Baltimore: Williams and Wilkins.
- Reynolds, Cecil R. and Brown, Robert T. (1984), eds. *Perspectives on bias in mental testing*. New York: Plenum Press.