

# **The Achievement Gap: Test Bias or School Structures?**

**National Association of Test Directors 2004 Symposia**

Organized by:

**Thel Kocher**  
Edina (MN) Schools

Edited by:

**Joseph O'Reilly**  
Mesa (AZ) Schools



This is the twentieth volume of the published symposia, papers and surveys of the National Association of Test Directors (NATD). This publication serves an essential mission of NATD - to promote discussion and debate on testing matters from both a theoretical and practical perspective. In the spirit of that mission, the views expressed in this volume are those of the authors and not NATD. The papers and discussant comments in this volume were presented at the April, 2004 meeting of the National Council on Measurement in Education (NCME) in San Diego.

The authors, organizer and editor of this volume are:

Thel Kocher, Organizer  
Edina Public Schools  
ISD 273 District Office  
Room 200  
5701 Normandale Road  
Edina, Minnesota 55424  
[thekocher@edina.k12.mn.us](mailto:thekocher@edina.k12.mn.us)

Steve Schellenberg  
Saint Paul Public Schools  
360 Colborne St.  
Saint Paul, MN 55102  
[steve.schellenberg@spps.org](mailto:steve.schellenberg@spps.org)

Stephen G. Sireci  
University of Massachusetts – Amherst  
Educational Policy, Research and Administration  
111 Infirmary Way Ofc 2  
Amherst, MA 01003-9329  
[sireci@acad.umass.edu](mailto:sireci@acad.umass.edu)

Margaret Jorgensen  
Harcourt Assessment  
19500 Bulverde Road  
San Antonio, Texas 78259-3701  
[margie\\_jorgensen@harcourt.com](mailto:margie_jorgensen@harcourt.com)

Jennifer McCreadie  
George Mason University  
College of Educ & Human Development  
A339 Robinson Hall  
4400 University Drive  
Fairfax, Virginia 22030-4444  
[jmccread@gmu.edu](mailto:jmccread@gmu.edu)

Glynn Ligon  
ESP Solutions Group  
1510 W. 34th Street  
Austin, TX 78703  
[Gligon@esp.com](mailto:Gligon@esp.com)

Joseph O'Reilly, Editor  
Mesa Public Schools  
63 East Main St #101  
Mesa AZ 85201  
[Joreilly@mpsaz.org](mailto:Joreilly@mpsaz.org)

## **Table of Contents**

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

**Steve Schellenberg.....1**

### **The Role of Sensitivity Review and Differential Item Functioning Analyses in Reducing the Achievement Gap**

**Stephen G. Sireci .....19**

### **A Test Publisher Perspective**

**Margaret Jorgensen.....41**

### **A District Perspective**

**Jennifer McCreadie .....46**

### **Discussant Comments**

**Glynn Ligon.....58**

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

**Stephen J. Schellenberg**

**Saint Paul (MN) Public Schools**

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.

# How Psychometricians Can Help Reduce the Achievement Gap: Or can they?

**Stephen G. Sireci**

**University of Massachusetts Amherst**

Standardized tests are used throughout educational systems in the United States and the No Child Left Behind Act of 2001 (NCLB, 2002) has ensured their place in this system for the foreseeable future. For decades, standardized tests in education have provided data that shows certain minority groups, such as African Americans and Hispanic/Latino(a)s, score significantly lower on these tests than non-minority groups such as Caucasians (U.S. Department of Education, 2002). This finding is known as the achievement gap, which manifests itself as a performance gap between minorities and non-minorities on educational achievement data.

The achievement gap is disturbing because it suggests inequities in our educational system and it illustrates that not only are many children in the system are being left behind, but also that such children are more likely to come from certain cultural groups. The causes of the achievement gap are difficult to isolate, although many

theories and strong political views have been put forward (e.g., Singham, 2003). In fact, standardized tests themselves are often cited as a contributing factor (e.g., English, 2002). In this paper, I do not address or investigate causes of the achievement gap. However, I do not think the gap can be blamed on standardized testing. To do so is like blaming the thermometer for a fever. Psychometricians and test developers have been aware of the achievement gap for decades and have been as troubled by it as anyone else. In fact, this awareness has had a significant impact on test construction and validation practices.

In this paper, I describe the test development and evaluation practices conducted by testing agencies to build tests that are as fair as possible to all examinees, including identifiable sub-groups such as cultural minorities. I draw from the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999) and other testing guidelines to show how these practices are universally endorsed by the measurement and educational research professions. I also discuss areas in which test development and validation practices could be improved to facilitate fairness in testing, and I present suggestions for further research into how test construction practices may affect performance differences between minority and non-minority test takers.

## **STANDARDS FOR FACILITATING EQUITY IN TESTING**

I have been a psychometrician for almost fifteen years. In that time, I have worked for several national testing companies, a local school district, and as an advisor to many licensure, college admissions, and statewide achievement testing programs. Although I certainly cannot defend all educational tests, I can assert with confidence that every testing program with which I have been associated strives to ensure fairness to all examinees. At the same time, I must admit that these efforts are essentially voluntary, although they may be partly driven by market and profit-oriented factors. The fact is,

there is no formal audit mechanism for the educational testing industry. That is, there is no Federal oversight organization like the Food and Drug Administration. The closest authoritative mechanism we have is a long history of professional guidelines currently embodied in the *Standards for Educational and Psychological Testing*<sup>1</sup> (AERA et al. 1999, hereafter referred to as the *Standards*). Testing agencies take these standards seriously and do their best to adhere to them. For one thing, it is well known that the courts consistently use these standards when tests are challenged through the legal system (Sireci & Green, 2000).

The *Standards* provide several guidelines for reducing or eliminating characteristics of tests that may adversely impact the test scores of minority groups. In fact, a chapter entitled “Fairness in Testing and Test Use” specifically addresses this topic. In that chapter, the *Standards* lament “Absolute fairness to every examinee is impossible to attain, if for no other reasons than the facts that tests have imperfect reliability and validity in any particular context...” (p. 73). The *Standards* also discuss different perspectives on test fairness. Using this discussion, they ultimately provide an operational definition of fairness as a situation where “examinees of equal standing with respect to the construct the test is intended to measure...on average earn the same test score irrespective of group membership” (p. 74). Furthermore, the *Standards* characterize fairness as a lack of bias and emphasize the importance of ruling out potential biasing factors in a test. This characterization is clear in their advice to testing agencies that “...consideration of bias is critical to sound testing practice” (p. 74).

The Fairness in Testing and Test Use chapter of the *Standards* advocate two test development practices that are designed to build fairness into the test development process. These two practices are *sensitivity review* and analysis of *differential item functioning*. In addition, the *Standards* describe important validation practices that are

---

<sup>1</sup> The first version of the *Standards* dates back to a document prepared by the American Psychological Association in 1952 (APA, 1952). The first joint AERA/APA/NCME version was published in 1954 (APA, 1954).

designed to evaluate test fairness post hoc. Two practices of particular importance to the evaluation of test bias are analysis of the consistency of test structure across different groups of test takers and the analysis of differential predictive validity. In the next sections of this paper, I describe these suggested practices, beginning with those related to test development (i.e., sensitivity review and differential item functioning).

## **FACILITATING FAIRNESS USING SENSITIVITY REVIEW**

Sensitivity review refers to the process of having a diverse group of professionals review tests to flag material that may unintentionally interact with demographic characteristics of some test takers. As Standard 7.4 describes

*Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain. (AERA et al. 1999, p. 82)*

The process of sensitivity review goes beyond identification and elimination of test content that may be construed as offensive (Ramsey, 1993; Sireci & Mullane, 1994). Such a review also seeks to identify test material that is irrelevant to the construct measured, but may provide an advantage or disadvantage to members of one or more sub-groups of examinees. That is, sensitivity reviews attempt to uncover unintended bias in tests. A classic example of construct-irrelevant bias creeping into an assessment is a math test with word problems coming entirely from an ethnocentric context such as baseball. For this reason, most test developers employ a panel of sensitivity reviewers to review preliminary test forms for material that may

- (a) be construed as offensive to particular groups of individuals,
- (b) portray groups of individuals unfavorably, or in a stereotypical fashion,
- (c) be advantageous to one group, and/or disadvantageous to another, or
- (d) be unfamiliar to certain groups of individuals.

Sensitivity review is often seen as a final check on potential bias in a test. Unlike a technical content review, the reviewers need not necessarily be experts in the subject

domain tested, although such expertise is desirable. Instead, the reviewers are selected for their knowledge of specific cultural groups and how test takers from such groups may interpret test material. Sensitivity reviewers are typically a diverse group consisting predominantly of minority group members. By reviewing tests for potential bias, sensitivity review can improve the content validity of a test, but it should not be confused with a content review. As Sireci and Mullane described

*“[test developers] strive to ensure that the content of their tests is representative of and relevant to the domain tested, and that trivial or superfluous material is not present. Therefore, it can be argued that sensitivity review is built into the test development process under the rubric of content validation. After all, if the entire content of a test has been demonstrated to be relevant to the content domain, there should be no material contained therein that is offensive or unfamiliar to examinees who are knowledgeable of the domain. This argument is flawed, however, because evaluations of test content focus on how well the items that comprise a test represent the content domain, rather than on evaluating the context within which the items are framed. (pp. 22-23)*

To provide an idea of the types of judgments made during a sensitivity review, two examples of test material that was flagged in a sensitivity review are presented in Figures 1 and 2 [These figures are not available for publication but were part of the presentation]. Figure 1 presents a political cartoon that was submitted as part of a set of items for a social studies test.

The content to be tested using the cartoon is knowledge of the breakup of the Soviet Union (it portrays a meeting of the “Amalgamated Evil Empires” with many empty seats and the representative from China saying to the leaders of Cuba, Viet Nam and North Korea “Move that we dispense with calling roll”). On one level, the cartoon is funny. It is also politically accurate, since Ronald Regan used the term “evil” to describe the communist block. However, for test takers whose cultural heritage originates from China, Cuba, North Korea, or Vietnam, the adjective “evil,” emblazoned at the top of the cartoon, may be offensive. Although such material may not in itself lead to lower test scores for test takers from these groups, it may bring forward emotions that are not conducive to optimal test performance.

This cartoon highlights both the importance of sensitivity review and its contentious nature. Many people believe that eliminating material such as this cartoon makes tests

less authentic and leads to essentially sterile test content. The important point of sensitivity review is to *flag* such test material so that the test developers and content experts can decide whether the context and content of the item is content-relevant or whether it is inappropriate. A common rule of thumb is that if the same test objective can be measured just as well in a more innocuous context, the item should be replaced (Ramsey, 1993; Sireci & Mullane, 1994).

The item in Figure 2 (a baseball themed test question) was not flagged for reason of offensiveness, but rather it contains construct-irrelevant material that would provide an advantage to some test takers and disadvantage others. To successfully answer this arithmetic item, one would need to know that there are nine innings in a baseball game. Therefore, the item is biased against anyone who does not know this fundamental aspect of our national pastime. Such an item would contribute to test bias and would inflate sub-group differences such as the achievement gap. This item would never make it through a legitimate sensitivity review process.

## **ANALYSIS OF DIFFERENTIAL ITEM FUNCTIONING**

The second guideline in the *Standards* that reflects an important test development step aimed at promoting fairness and reducing the achievement gap is analysis of differential item functioning (DIF). DIF refers to a situation where test takers who are considered to be of equal proficiency on the construct measured, but who come from different groups, have a different probability of earning a particular score on a test item.

To understand DIF, three related concepts must be distinguished: item impact, DIF, and item bias. Item impact refers to a significant group difference on an item, for example when one group has a higher proportion of test takers answering an item correctly than another group. Item impact may be due to true group differences in academic performance or due to item bias. DIF is a statistical observation that

involves *matching* test takers from different groups on the characteristic measured and then looking for performance differences on an item. Test takers of equal proficiency who belong to different groups should respond similarly to a given test item. If they do not, the item is said to function differently across groups and is classified as a DIF item (see Clauser & Mazor, 1998, or Holland & Wainer, 1993 for more complete descriptions of DIF theory and methodology). Item bias is present when an item has been statistically flagged for DIF *and the reason for the DIF is traced to a factor irrelevant to the construct the test is intended to measure*. Therefore, for item bias to exist, a characteristic of the item that is unfair to one or more groups must be identified. Thus, a determination of item bias requires subjective judgment that a statistical observation (i.e., DIF) is due to some aspect of an item that is irrelevant to the construct measured. That is, difference observed across groups in performance on an item is due to something unfair about the item.

If an item were flagged for DIF, it does not automatically mean that the item is biased. Item bias requires additional subjective judgment that the reason for DIF is construct-irrelevant. Thus, DIF is a necessary, but insufficient condition for item bias.

An example of an item flagged for DIF is presented in Figure 3 (not available). This item comes from analysis of an eighth-grade English Language Arts test from the Massachusetts Comprehensive Assessment System. Below the text for the item is a conditional p-value plot, which displays the proportion of Black and White eighth-graders who answered the item correctly, conditional on total test score. This plot illustrates the “matching” aspect of DIF analyses because the plotted points on each line represent test takers in each group who earned similar scores on the test. In this figure, it is evident that the Black students were more likely to answer the item correctly than the White students throughout the bulk of the test score distribution. An analysis of the content of the item reveals that the vocabulary word tested, “ebony,” is likely to be more familiar to African American students than to Euro-American students. Ebony is a dark-colored wood and it is also the name of a popular magazine targeted to African-Americans. Thus, it is not surprising Black test takers were more

likely to know the correct synonym for this term. But is this item biased? Differential cultural familiarity does not necessarily signify bias. In this case, the term was considered part of the curriculum tested and was not removed from the test.

The *Standards* do not go so far as to require testing agencies to conduct DIF analyses, but they do suggest that the results from such analyses be thoroughly investigated. As Standard 7.3 asserts

*When credible research reports that [DIF] exists across age, gender, racial/ethnic, cultural, disability and/or linguistic groups in the population of test takers...test developers should conduct appropriate studies where feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups. (AERA et al. 1999, p. 81)*

With respect to the achievement gap, it is presumed that if items that are unfair to minority test takers can be identified and screened out, the performance differences noted between non-minority and minority groups will diminish. Most developers of K-12 tests use DIF analyses at the item pretest stage and use DIF results to screen out potentially problematic items for inclusion in a test. DIF analyses are also conducted after tests are administered to ensure no item bias is present.

It is somewhat perplexing that the *Standards* do not explicitly encourage test developers to perform DIF analyses as a regular part of the test development process. Perhaps the authors had only the most conscientious testing agencies in mind when they wrote “Although DIF procedures may hold some promise for improving test quality, there has been little progress identifying the causes or substantive themes that characterize items exhibiting DIF” (p. 78). It is true many items that are flagged statistically for DIF are never considered biased because no reason or theory can be brought forward to explain the DIF. However, there are many instances where items were flagged for DIF and the cause was interpretable by content specialists (Allalouf, Hambleton, & Sireci, 1999). In some cases, the cause was seen as construct-irrelevant and an item was removed from a test for reasons of bias (e.g., Schmitt & Bleistein, 1987). In other cases, the cause of the DIF was considered construct-relevant, and the item was not excluded from the test.

Before leaving the topic of DIF, many psychometricians have correctly pointed out that analysis of bias at the item level is insufficient for evaluating test fairness for a couple of reasons. First, DIF analyses typically use total test score as the criterion for matching students across groups. If there is systematic bias in the total test score, the matching criterion is biased and the detection of DIF will be flawed. For this reason, analysis of test bias, described in the next section, is necessary for a thorough evaluation of test fairness. A second reason why item-level analyses are insufficient for analyzing bias is that very small levels of DIF that may not reach statistical significance may have an aggregate effect across a set of test items (Shealy & Stout, 1993; Wainer, Sireci, & Thissen, 1991). That is, if there were a few items that functioned differentially against African Americans, even though each item may not exhibit a large enough difference to be flagged, adding the differences across all items may have a negative effect on total test performance. A related possibility is that DIF items may balance each other out. For example, there may be one item on a test that exhibits DIF against African Americans and another item that exhibits DIF against Euro-Americans. Some researchers argue that including both items is fair and leads to less sterile tests (Humphreys, 1970). Although this idea is controversial, most psychometricians agree that DIF aggregation and cancellation should be investigated.

## **OTHER STATISTICAL ANALYSES FOR EVALUATING THE PRESENCE OF BIAS**

The *Standards* and most psychometric textbooks describe two other common statistical approaches for evaluating tests for potential biasing factors that could contribute to an observed achievement gap between groups. One of these approaches is an internal analysis that focuses entirely on test takers' responses to items. The other approach is external and relates test scores to other criteria. The internal analysis is typically described as a comparison of the factor structure of a test across groups.

The external analysis is typically described as an analysis of differential predictive validity.

### **Analysis of Test Structure**

The internal “structure” of a test is typically described by determining the most parsimonious set of factors that can be used to summarize test takers’ responses to test items. Procedures like factor analysis and multidimensional scaling can be used to identify common clusters of items that may explain different patterns of examinee performance. Many educational tests are designed to be unidimensional, which means only one factor is presumed to underlie examinees’ response data. With respect to test bias, analysis of the consistency of the factor structure across test data obtained from minority and non-minority examinees is relevant. The *Standards* explain why comparison of test structure across groups of examinees is relevant to test fairness:

*...evidence of bias related to response processes may be provided by comparisons of the internal structure of the test responses for different groups of examinees...In situations where internal test structure varies markedly across ethnically diverse cultures, it may be inappropriate to make direct comparisons of scores of members of these different cultural groups. (AERA et al., 1999, p. 78)*

In less esoteric terms, if the internal structure of a test differs across groups, it may signify that the average strengths and weaknesses of test takers are not consistent across groups. Such a finding is strange for a test that is designed to test the same construct for all groups. Hence, differential test structure may suggest construct nonequivalence across groups. That is, the test may be measuring something different in one group, relative to another.

Analysis of test structure across groups is typically not done before operational testing because of the data requirements for such analyses. However, it is rare that such analyses reveal “markedly” different structures (cf. Robin, Sireci, & Hambleton, 2003). This observation may be due in part to successful screening of DIF prior to the assembly of test forms. It may also be due to the fact that the constructs tested on educational tests are consistent across diverse groups due to similar curricula and educational objectives.

## Differential Predictive Validity

The second popular statistical technique for evaluating test fairness is differential predictive validity (DPV). *Predictive validity* is the degree to which test scores accurately predict scores on a criterion measure (Wainer & Sireci, in press). DPV investigates whether the relationship between test and criterion scores is consistent across examinees from different groups. To conduct an analysis of DPV, data on a relevant external criterion must be available for both minority and non-minority test takers. For this reason, most studies of DPV have been conducted on admissions tests, with grade-point-average (GPA) as the validation criterion.

Multiple regression analyses are typically used to evaluate DPV. The simplest form of a regression model used in this context is

$$y = b_1X_1 + a + e$$

where  $y$  is the predicted criterion value (e.g., GPA),  $b_1$  is a coefficient describing the utility of a predictor (e.g., test score) for predicting the criterion,  $a$  is the intercept of the regression line (i.e., the predicted value of the criterion when the value of the predictor is zero), and  $e$  represents error (i.e., variation in the criterion not explained by the predictors). The residuals ( $e$  in the equation) represent the difference between the criterion value predicted by the equation and the actual criterion score. DPV can be evaluated by fitting separate regression lines for each group, and then testing for statistically significant differences between the slopes and intercepts (Linn, 1984; Pedhazur & Schmelkin, 1991; Wainer & Sireci, in press). Such an approach is preferable when there are sufficient data for each group of interest (minimally, 100 students, preferably more). When sample sizes do not permit fitting separate regression equations, a single equation is fitted to the data for all examinees, and the residuals are analyzed for patterns of *over-prediction* (predicted scores are higher than criterion scores, i.e.,  $Y - \hat{Y} < 0$ ) or *under-prediction* (predicted scores are lower than

criterion scores, i.e.,  $Y - \hat{Y} > 0$ ; Braun, Ragosta, & Kaplan, 1986; Koenig, Sireci, & Wiley, 1998).

Most studies of DPV have not found instances of test bias. In fact, where differences in predictive accuracy are found, the typical finding is that the test scores for minority examinees tend to over-predict their criterion scores. The lack of bias found in DPV studies could be due to the fact that it is the most conscientious testing agencies that conduct such analyses, and they are the least likely to produce tests that contain bias. Nevertheless, the important point to bear in mind is that the field of psychometrics has produced the statistical machinery to evaluate whether the predictive utility of test scores is compromised for specific sub-groups of examinees.

## **SUMMARY OF WHAT PSYCHOMETRICS HAS CONTRIBUTED TO FAIRNESS IN TESTING**

Standardized testing is designed to provide a level playing field for all test takers by making the test content, administration conditions, and scoring uniform for all test takers. Given the diversity inherent in typical populations of test takers, professional standards of test development require qualitative and statistical screening of test items and test forms, as well as comprehensive post hoc analyses of the relationship of test scores to other variables that can attest to the validity of the inferences that are made on the basis of test scores. Such test development and evaluation guidelines do not guarantee tests that are fair to all test takers, and they do not guarantee that characteristics of tests do not contribute to the achievement gap. However, they demonstrate what the psychometric community has done to minimize bias in testing.

## **CAN PSYCHOMETRICIANS DO MORE TO INCREASE FAIRNESS IN TESTING?**

Up to this point, I have described the steps and statistical procedures psychometricians use to promote fairness in testing. However, a review of these practices and the current state of affairs in testing allows for the exploration of other activities that could be undertaken to further facilitate fairness in testing.

### **Underrepresentation of Minorities in Psychometrics**

Although virtually all minority students in the U.S. take standardized tests, there are very few educational measurement professionals who are minorities, particularly African Americans (Sireci, 2000). In fact, only two of 15 members of The Joint Committee on Standards for Educational and Psychological Testing, who essentially wrote the *Standards* were African American, and only one was Hispanic/Latino<sup>2</sup>.

Furthermore, in a recent survey of members of NCME, Sireci and Khaliq (2002) reported that some minority members stated that many minorities felt the testing profession represented primarily “White” culture and value systems. Although this is certainly not the intent of testing programs (as all the aforementioned standards and practices attest), the lack of minorities in our field may help promote such an impression. To carry this thought further, test items are probably written primarily by non-minorities, edited by non-minorities, and evaluated statistically by non-minorities. If this situation were true, sensitivity review is the only area in which minority involvement explicitly occurs.

Given the possibility of non-minority ethnocentrism inherent in educational assessment, I propose the following study to determine if this ethnocentrism contributes to the achievement gap. The study would involve recruiting qualified

---

<sup>2</sup> It should be noted, however, that this Committee did solicit comments from minority interest groups including the Equal Employment Advisory Council, International Association for Cross-Cultural Psychology, NAACP Legal Defense and Educational Fund, Society for the Psychological Study of Ethnic Minority Issues, and the U.S. Equal Employment Opportunity Commission. Nevertheless, including members from one or more of these groups on the Committee, may be helpful for ensuring minority issues are fully addressed in the *Standards*.

subject matter experts to write items for a particular educational achievement test. Equal numbers of African American and Euro-American item writers would be recruited. The qualifications for recruiting members in each group would be the same, and all item writers would receive the same item writing training. All items would be edited for inclusion on the test using a diverse team of item editors—with equal representation of African American and Euro-American item editors. Two versions of the achievement test would be created. One version would select items from the pool written by African American item writers, the other from the pool of items written by Euro-American item writers. The tests would then be spiraled and randomly administered to sufficient numbers of African American and Euro-American students for whom the test was intended. If the African American students do better on the “Black” version of the test, relative to their White peers, than they do on the “White” version (again relative to their White peers), then it could be that current test development procedures are too White-centric. If the results show no differences, the claim that ethnocentrism in test development contributes to the achievement gap would not be supported.

### **Re-organizing Testing Standards to Focus on Universal Test Design**

As mentioned earlier, the *Standards* contain a chapter devoted to the topic of fairness in testing. Although inclusion of this chapter is certainly very positive, it is surprising that explicit consideration of issues of fairness is not found in “Part 1” of the document, particularly in the chapter “Test Development and Revision.” Including standards for sensitivity review and DIF in that chapter, and then reiterating them in the fairness chapter, would send a stronger message to test developers that the diversity of the test taker population must be considered from the outset.

Recently, the concept of *universal test design* (UTD) has been introduced to suggest the development of tests that are more attuned to the differing needs of sub-groups of examinees (Thompson, Blount, & Thurlow, 2002). Essentially, this concept calls for test construction practices focused on eliminating construct-irrelevant variance and

more flexible test administration conditions (e.g., elimination of time limits). It would be useful if the next revision of the *Standards* incorporated this concept into the chapter on test development. Comparisons of the achievement gap resulting from tests developed using UTD principles with those not using UTD principles may provide interesting information on the degree to which test development model interacts with group differences in test performance.

A related issue is evaluating the degree to which testing programs actually adhere to the *Standards*, particularly with respect to standards related to issues of fairness. As mentioned earlier, there is no Federal audit mechanism for educational testing companies. However, some non-profit organizations, such as the Buros Center for Testing and Educational Testing Service, will provide a service to testing companies to let them know how well their practices adhere to professional standards. It would be good if such audits were not voluntary. In fact, demonstration of adherence to professional standards of test development should have been written into the NCLB legislation. The Commission on Instructionally Supportive Assessment (2001) encouraged states to ensure that their testing contractors adhered to the *Standards*. An important question we should ask all test developers is “Do you read the *Standards* and then develop tests accordingly, or do you develop tests first and then see how well they meet the *Standards*? My suspicion is that many test developers do the latter rather than the former.

### **Require More Studies of Content Validity**

Content validity is the degree to which an assessment represents the content domain it is designed to measure (Sireci, 2003). When an assessment is judged to have high content validity, the content of the assessment is considered to be congruent with the testing purpose and with prevailing notions of the subject matter tested. In such situations, there is less opportunity for biasing factors to enter into a test. Therefore, empirical studies of content validity can facilitate test fairness, which in turn lead to a reduction in the achievement gap.

Unfortunately, empirical studies of content validity are rare. Many test developers have several iterations of technical and content reviews of items, but *independent* analysis of the degree to which items measure their intended objectives are hardly ever conducted (Sireci, 1998). Given that content validity is prerequisite to the validity of scores from educational tests, the lack of empirical research in this area is troubling. By making independent evaluations of test content routine, the fairness of educational tests to all test takers is likely to increase.

## **SUMMARY AND CONCLUSIONS**

In this paper, I described test development and evaluation activities that are supported by professional standards in the measurement community and that are conducted by many testing agencies. The activities described include sensitivity review, analysis of DIF, comparison of test structure across different groups of examinees, and evaluation of DPV. These four activities are designed to facilitate fairness in testing and consequently ensure that test limitations do not contribute to undesirable outcomes such as the achievement gap. I also provided several suggestions for testing practice and for further research that could provide more information regarding the degree to which testing characteristics may contribute to the achievement gap. These suggestions were:

1. Conduct experimental studies to evaluate the effect of culture of the test developers on the minority test takers' test performance.
2. Evaluate the degree to which tests developed using UTD may reduce achievement gaps.
3. Encourage testing agencies to routinely conduct content validity studies using independent subject matter experts.
4. Encourage AERA, APA, and NCME to describe the importance of sensitivity review and DIF analyses in the chapter on test development in the next

revision of the *Standards*. In addition, studies of differential factor structure and DPV should be mentioned in the validity chapter.

5. Encourage AERA, APA, and NCME to contain better minority representation on the Joint Committee that authors the next version of the *Standards*.

My discussion of the benefits of quality test development and evaluation practices does not ensure that test characteristics do not contribute to the achievement gap. However, I hope it illustrates the great concern that psychometricians have for equity issues, and the efforts they go through to minimize potential biasing factors. I also hope some of the suggestions outlined in the paper may prove fruitful for better understanding the interaction between cultural group membership and test performance, and how psychometric research and practice can contribute to the elimination of the achievement gap.

## REFERENCES

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the sources of differential item functioning in translated verbal items. *Journal of Educational Measurement*, *36*, 185-198.

American Psychological Association, Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist*, *7*, 461-465.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*, (2, supplement).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Braun, H., Ragosta, M., & Kaplan, B. (1986). The predictive validity of the Scholastic Aptitude Test for disabled students (*Research Report 86-38*). New York: College Entrance Examination Board.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice, 17(1), 31-44.

Commission on Instructionally Supportive Assessment, (2001, October). Building tests to support instruction and accountability: A guide for policymakers. Available at [www.asa.org](http://www.asa.org), [www.naesp.org](http://www.naesp.org), [www.principals.org](http://www.principals.org), [www.nea.org](http://www.nea.org), and [www.nmsa.org](http://www.nmsa.org). Author.

English, R. W. (2002) On the intractability of the achievement gap in urban schools and the discursive practice of continuing racial discrimination. *Education and Urban Society*, 34, 298-311.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum.

Humphreys, L. R. (1970). A skeptical look at the pure factor test. In C. E. Lunneborg (Ed.). *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp 23-32). Seattle: University of Washington.

Koenig, J. A., Sireci, S. G., & Wiley, A. (1998). Evaluating the predictive validity of MCAT scores across diverse applicant groups. *Academic Medicine*, 73, 65-76.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement*, 21, 33-47.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P.W. Holland & H.Wainer (Eds.), Differential item functioning (pp. 367-388). Hillsdale, New Jersey: Lawrence Erlbaum.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, 3, 1-20.

Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning of black examinees on Scholastic Aptitude Test analogy items* (Research Report 87-23). Princeton, NJ: Educational Testing Service.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58, 159-194.

Singham, M. (2003). The achievement gap: Myths and reality. *Phi Delta Kappan*, 84, 586-571.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83-117.

Sireci, S. G. (2000). Recruiting the next generation of measurement professionals. *Educational Measurement: Issues and Practice*, 19(4), 5-9.

Sireci, S. G. (2003). Content validity. *Encyclopedia of psychological assessment* (pp. 1075-1077). London: Sage.

Sireci, S. G., & Green, P. C. (2000). Legal and psychometric criteria for evaluating teacher certification tests. *Educational Measurement: Issues and Practice*, 19(1), 22-31, 34.

Sireci, S. G., & Khaliq, S. N. (2002). NCME members' suggestions for recruiting new measurement professionals. *Educational Measurement: Issues and Practice*, 21(3), 19-24.

Sireci, S. G., & Mullane, L. A. (1994). Evaluating test fairness in licensure testing: The sensitivity review process. *CLEAR Exam Review*, 5(2), 22-28.

Thompson, S., Blount, A., & Thurlow, M. (2002). A summary of research on the effects of test accommodations: 1999 through 2001 (*Technical Report 34*). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 2003, from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm>

U.S. Department of Education, National Center for Education Statistics (2002). *The condition of Education 2002, NCES 2002-025*, Washington, DC: U.S. Government Printing Office.

Wainer, H., & Sireci, S. G. (in press). Item and test bias. *Encyclopedia of social measurement*. San Diego: Elsevier.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, 28, 197-219.

# **The Achievement Gap: Test Bias or School Structures? A District Perspective**

**Jennifer McCreddie**

**Indianapolis Public Schools**

The Achievement Gap: Test Bias or School Structures? From a district perspective, my answer to that question is yes, both, but also so much more. It's hard for me to see it as a simple question of test bias or school structures. The achievement gap is the result of complex social issues that are far broader. Poverty and segregation play critical roles in the current achievement gap, and they have a long history.

From a district perspective, our charge is to educate all of our students. We must do this despite the fact that many of our children come to us facing serious difficulties in their lives. They may be angry and frustrated, witnesses to crime and drug use, victims of health problems, homelessness, hopelessness, hunger. Schools are expected to address these cultural, economic, social, health and safety, nutrition needs, as well as

change low expectations, lack of aspiration, loss of hope. Schools must meet all the needs of their students, to make them ready to learn and, then, to teach them. Schools have to go beyond what has been historically true about student achievement gaps to ensure that the conditions that led to those gaps will not also predict their continuation into the future. Is it any wonder that the schools sometimes fail? Or, should I say "need improvement?"

In addressing the achievement gap, we should acknowledge the existence of several achievement gaps. The first gap we think of in the context of *Brown vs. Topeka Board of Education*, is typically racial/ethnic - between white and Asian students on the one hand, and African-American and Latino students on the other. However, there are more gaps - socioeconomic, which overlaps the other categories, gaps between English language learners and native English speakers, between students with disabilities and without and, still, between genders.

To answer the question at hand, both test bias and school structures certainly can be addressed and improved. All tests are biased in some way and to some extent. Test bias is routinely addressed through both examination and correction for item bias and sensitivity review. The nature and modes of assessment tasks should also be considered, but practical limitations such as cost will often dictate the choice. Steve Sirici and Margie Jorgensen discuss here many ways in which test bias is being managed and minimized.

Steve Schellenberg wrote, "We must not resort to the glib mantra that "It's not the test, but the use of the results, that is biased." I take up this argument, not at all glibly, because I believe that the use of test results has been taken to an extreme with the ESEA reauthorization of 2001, better known as No Child Left Behind.

NCLB is civil rights legislation with a laudable goal. At its best, its illumination of the differences in performance among various groups of children ultimately spotlights the inequities and inequalities that persist in our society and in our schools today. However, the test-driven nature of this accountability system and its high stakes for

schools, districts, and their staffs are problematic. The intent of the law may be distorted by an overemphasis on numbers. The effect of the law is to blame and punish the children who most need help and the schools that serve them.

I will address three points about NCLB before returning to a discussion of school structures, where school districts have to focus our energy and attention:

- The target of 100 percent of our students achieving proficiency is unrealistic. It is our job, our teachers' job, to have the highest expectations for all students and to challenge, lead, push every student to learn. We must also help some students and their families to have higher expectations and aspirations for themselves. However, unless "proficient" is defined at an unacceptably low level, 100 percent of students will not achieve it. That should never mean that we don't work to achieve that level and it should never mean that we decide in advance who will not achieve it.
- The "all or nothing" nature of NCLB is demoralizing to educators. There are so many ways to fail and only one way to succeed. If any subgroup misses the target on any of the factors, the school or district has not achieved adequate yearly progress (AYP); but only when all subgroups meet the target is AYP achieved.
- The burden of NCLB could ultimately fall only on those schools and districts with populations large and diverse enough to have multiple subgroups. As revisions create flexibility for groups or factors that affect even relatively homogeneous and affluent schools and districts, such as students with disabilities or the percentage of students participating in the assessment,

only those with racial/ethnic and socioeconomic diversity will face consequences.

Will enough voices remain to protest the unfairness and to advocate for all of our students, especially for those who need it most, once adjustments are made to protect the most homogeneous and affluent schools, districts, and states?

Back to the achievement gap, because we must continue to attack it within the larger context - there are some rays of hope:

- The Council of the Great City Schools published Beating the Odds IV in March of this year. The Council is an organization representing over 60 of the country's 100 largest school districts. Achievement test data from member districts show progress - students are achieving at higher rates, and urban districts are improving at faster rates than their states, which is good news.
- In my district of over 40,000 students, only one subgroup achieved the target on all factors. However, our third graders overall in fall 2003 performed better than did their counterparts in 2002, as did every subgroup. At the higher grades, most subgroups performed better in 2003 than did those in 2002. But the gaps did not narrow - at least not yet.

What I fear will narrow is the curriculum! Not the formal, written curriculum, of course, but what is actually being taught. To the extent that test scores and numbers are being emphasized instead of students and their learning, we risk seeing teaching to the test instead of to the standards, and teaching facts and skills in isolation - something we know isn't very effective.

Still, after all this, education remains the best hope for overcoming historical inequities. We turn again to look at school structures - curriculum, instruction, assessment, students, teachers, and schools. We continue to look for new, better ways to do more, more effectively, to help our students to learn. Schools are attempting to improve, to reform, to transform themselves both with and without the resources to do so. (And there are rarely enough resources, whether we are talking about money, people or time.)

What do we need to improve about school structures to reduce the achievement gaps?

- We need to have a clear direction and focus on improving student learning in our schools. We need to have a sense of urgency because the needs of our students are urgent, not because the numbers may look bad.
- We need to recruit, retain, and continue to develop good, effective teachers (highly qualified teachers, if you will). We need to find ways to match the most effective teachers with the students most at risk, rather than placing the newest or least effective teachers at schools with the highest poverty and greatest needs.
- We need teachers who are licensed and experienced in the areas they teach, especially in math, science, and special education.
- We need to provide, promote, and enable appropriate and effective professional development.
- We need to give teachers time to think, to work and plan together, to talk about important issues in their work.
- We need to fully implement carefully selected initiatives or programs.

- We need to evaluate carefully and thoughtfully instructional materials in relation to their focus, their match to the curriculum and to children's needs. The chosen materials must then be integrated into a coherent instructional program, not just tagged on.
- We need to use data of all kinds to help us know what's making a difference for whom. We need to rely less solely on tests and more on multiple types and sources of data.

I already mentioned the high expectations we must have for all students. We need to have a rigorous curriculum for virtually all students, and we need to be sure that they are learning in the early years so that they have the foundation to complete - successfully - advanced courses. We want them to expand their options during high school so that they have choices when they graduate. We need to find ways for our students to have opportunities and aspirations beyond their life experience so that their next steps after high school take them far beyond where they live and what they know now.

# The Achievement Gap: Test Bias or Real Differences? A Test Publisher's Perspective

Margaret A. Jorgensen

Harcourt Assessment, Inc.

The achievement gap in the United States is a disturbing symptom of an epidemic ailment in our educational system. For all the reasons cited by both Schellenberg (2004) and Sireci (2004), professionals who are involved daily in the development of achievement tests and the use of the resulting data must probe deeply to answer the question: *Are the gaps the result of true differences in achievement or are they the result of bias in the measurement instruments?* (Schellenberg, p. 2).

Note: This paper is Copyright © 2004 by Harcourt Assessment, Inc. All rights reserved.

This paper describes the development system used by a test publisher to ensure that bias in the assessment instrument has been minimized given our current understanding

of the measurement of achievement. We must continue, however, to have a broader, more sensitive discussion of what must be known about teaching, learning, and assessment before the answer to Schellenberg's question can be known.

## **IDENTIFYING THE DIFFERENCES**

Test developers, defined in the broadest sense, have a professional and ethical responsibility to remove barriers to content access for any individual. The goal is to publish a test that identifies achievement differences only, and not differences reflecting culture, ethnicity, geography, socioeconomic status, etc. The general practices, techniques, and methodologies used to reach this test publishing goal are well- documented and widely accepted.

Deeply underpinning this discussion is the fundamental issue of test purpose. Consider, for example, a norm-referenced test that does not reveal differences among a group of students. Likewise, consider a standards-based assessment where every student is "proficient." As a test publisher, I would suggest that to administer either of these assessments is a waste of time and is probably an expense that should not be allowed.

If we truly believe and have evidence to support our belief that there are no differences among students assessed using a norm-referenced or standards-based interpretive framework, there is no legitimate reason to build and administer a test that confirms sameness. Therefore, the clear purpose of assessment is to identify differences in student achievement. Only by understanding those differences can we improve instruction for each student.

For educators (again, defined in the broadest sense), there is less insight to be gained from sameness than there is from identifying differences. There is also the intuitive recognition that if a measured variable indicates that we are all exactly the same, the

variable itself is probably too broadly defined. If tests are expected to detect differences in the measured variable (e.g., achievement in a particular content area), how do test developers ensure that bias is minimized and only achievement differences are measured? The minimization of bias must be considered at all stages of the test development process: content, item construction, test format, test structure, administration directions, scoring, and score reporting.

For a test publisher developing tests for a wide range of customers, there are well-accepted and well-documented procedures. Some of them are subjective, such as having trained assessment developers making the right decisions about the content to be included. Some procedures involve sequential reviews of the content during the development process. A thorough content review process entails having representatives of various geographic, educational, social, and ethnic groups participate in the repeated re-examination of content to ensure that no word, graphic, content presentation, or other attribute of the content is biased for or against any test-taker. An example of a checklist typically used to guide content reviews follows:

Does the item provide access for the greatest number of test-takers? Is the item free from bias in the areas of:

<b>Item Bias and Sensitivity Issues</b>
Does the item provide access for the greatest number of test-takers?
Is the item free from bias in the areas of:
Gender?
Race?
Religion?
Socio-economic status?
Age?
Culture?
Is the item sensitive to:
Special -needs groups, such as physically disabled, visually impaired, and deaf and hard of hearing people?

Second-language learners?
Does the item avoid offensive or disturbing information?

Other methods used to eliminate bias are similar to those currently presented by Sireci (2004). These methods require sufficient representative sample sizes, appropriate analytic tools, expertise in interpreting analyses, and strategies for making adjustments based on these analyses while sustaining the purpose of the assessment.

For proprietary products developed by test publishers, the procedures and tools used to minimize bias reflect the best practices articulated in the scientific literature as well as the expertise and experience of professionals. Customers are increasingly sophisticated and knowledgeable and their response to the final product is, of course, the ultimate test of acceptability.

For custom-developed (work-for-hire) assessment products, the customer is often very involved in the subjective review process. This review is often public and the subject of open discourse. Politicians, media representatives, teachers, and education agency staff all participate in content review at some point as the test is developed and administered. The empirical processes for detecting bias are much less public and likely the result of limited knowledge.

From the subjective review of content to the empirical review of differences in item and test performance across groups, the processes for measuring achievement differences are established and accepted. However, these processes do not take us where we need to be in terms of *understanding* the differences. Schellenberg's question still remains unanswered.

## **UNDERSTANDING THE DIFFERENCES**

If best practice methods for identifying bias still cannot address the question of whether achievement gaps are the result of achievement differences or test bias, what is to be done? How can test publishers gain a deeper understanding of the root causes of achievement differences? I would like to suggest several possibilities.

What if items were constructed differently than they are for today's high stakes and norm-referenced tests? What if items were constructed so that the given answer could be interpreted against a knowledge hierarchy? Let me be specific by focusing on one type of item—the multiple-choice question.

Multiple-choice test items have a long history of helping instructors and policy makers understand what students know and can do. The incorrect answer options, or distractors, have been far less important in writing and editing items than the correct answer option has been. Ideally, distractors are structured to reflect typical student errors. However, information about those errors and what they reveal about student cognition has generally not been collected and analyzed. If the purpose of collecting information is to improve student understanding and to target instruction toward correcting errors in students' thinking, it makes sense to not only collect that information, but to structure items so the distractors yield more information about student errors in cognition and in understanding.

To differentiate between the distractor options, it is important for the types and levels of errors they reflect to be clearly stated and distinguishable from each other. This gives the item the power to differentiate among students choosing either distractor. This described function of the incorrect answer option is referred to as the *distractor rationale*. Using this type of item, a pattern of misconception or weakness may be revealed in a student's answer choices, thereby allowing an instructor to intervene with targeted instruction for that student.

Because an item is written to assess a particular learning standard associated with tasks ranging from simple to complex, there will also be a range of the types of

distractor rationales. Very simple Cognitive Level items that assess student recall may have little distinction among the incorrect answer options. For more complex items, however, distractor options that reflect a range of cognition levels and errors are possible.

The table below presents a hierarchy of the four levels of cognition that guide the writing of distractor options and rationales for reading items.

<b>Cognitive Level</b>	<b>Student Error</b>
<b>Level 1</b>	Makes errors that reflect focus on decoding and retrieving facts or details that are not necessarily related to the text or item. Student invokes prior knowledge related to the general topic of the passage, but response is not text based. These errors indicate that the student is grabbing bits and pieces of the text as he or she understands them, but the pieces are unrelated to the information required by the question being asked.
<b>Level 2</b>	Makes errors that reflect initial understanding of facts or details in the text, but inability to relate them to each other or apply them to come to even a weak conclusion of inference.
<b>Level 3</b>	Makes errors that reflect analysis and interpretation, but conclusions or inferences arrived at are secondary or weaker than ones required for correct response.
<b>Level 4</b>	Correct response.

Examples of items written using distractor rationales and the hierarchy of cognition levels follow:

Objective: Vocabulary

Hierarchically constructed item:

**Read this sentence from the story "Frogs and Toads."**

Both frogs and toads have a tail at first that disappears when they get older.

**What word has the same meaning as disappears as it is used in this sentence?**

A **disagrees** [Cognitive Level 1: look-alike word]

B **vanishes** [Cognitive Level 4: correct answer]

C **can be seen** [Cognitive Level 2: antonym]

D **becomes small** [Cognitive Level 3: related to the meaning, but not precise]

The same item structured in a traditional high-stakes way:

**Read this sentence from the story "Frogs and Toads."**

Both frogs and toads have a tail at first that disappears when they get older.

**What word has the same meaning as disappears as it is used in this sentence?**

A **turns green**

B **vanishes\***

C **jumps**

D **breaths**

all distractors are related to frogs, not to the meaning of the word "disappears."

Objective: Main Idea

Hierarchically constructed item:

**What is the main idea of the story "Frogs and Toads"?**

- A Frogs and toads share many differences and similarities.** [Cognitive Level 4: correct answer]
- B Frogs and toads are cute.** [Cognitive Level 1: prior knowledge, not text-based.]
- C Toads have shorter legs than frogs have.** [Cognitive Level 2: text-based detail unrelated to main idea]
- D Frogs are different than toads.** [Cognitive Level 3: only part of the time]

The same item structured in a traditional high-stakes way:

**What is the main idea of the story "Frogs and Toads"?**

- A Frogs and toads share many differences and similarities.\***
- B Frogs live closer to water than toads.**
- C Frogs and toads are like cousins.**
- D Frogs are different than toads.**


All distractors are essentially cognitive level 3: They are all related to the main idea, but are not the *best* answer.

Objective 3: Identifying Conflict

Hierarchically constructed item:

<p><b>What is the hare's main problem in "The Tortoise and the Hare"?</b></p> <p><b>A He does not like the tortoise.</b> [Cognitive Level 1: based on title]</p> <p><b>B He wants to run faster than the owl.</b> [Cognitive Level 2: incorrect character]</p> <p><b>C He loses the race even though he is fast.</b> [Cognitive Level 3: summary]</p> <p><b>D He is sure he will win so he stops trying.</b> [Cognitive Level 4: correct]</p>
---

The same item structured in a traditional high-stakes way:

<p><b>What is the hare's main problem in "The Tortoise and the Hare"?</b></p> <p><b>A He is lazier than the hare.</b></p> <p><b>B He falls asleep during the race.</b></p> <p><b>C He loses the race even though he is fast.</b></p> <p><b>D He is sure he will win so he stops trying.*</b></p>	 <table border="1"><tr><td><p>All distractors are essentially cognitive level 3: They are all related to the main problem, but are not the <i>best</i> answer.</p></td></tr></table>	<p>All distractors are essentially cognitive level 3: They are all related to the main problem, but are not the <i>best</i> answer.</p>
<p>All distractors are essentially cognitive level 3: They are all related to the main problem, but are not the <i>best</i> answer.</p>		

## CONCLUSION

By building assessments that provide increasingly precise information about why students choose the wrong answer, perhaps test publishers can answer Schellenberg's question about the cause of achievement gaps and help forge the link between assessment and instruction. If the differences are a function of learning and if our

society hopes to eliminate differences in levels of achievement, it seems reasonable to probe deeply into learning. Validity can no longer be all about measuring the right thing in the right way; that only focuses on getting to the correct answer. Validity must also be about (a) identifying how a student arrives at a wrong answer and (b) using that information to improve learning. Our focus must shift from dichotomously dividing students into those who have learned and those who have not. We must begin a profound exploration of the learning paths of students who are not learning what we expect them to know. It is only then that we can begin to fully understand the real differences in student learning.

## REFERENCES

Schellenberg, S. J. (2004, April). *Test bias or cultural bias: Have we really learned anything?* Paper presented at the AERA Annual Conference, San Diego, CA.

Sireci, S. G. (2004, April). *How psychometricians can help reduce the achievement gap: Or can they?* Paper presented at the AERA Annual Conference, San Diego, CA.

King, K. V. (2004). *Distractor rationales*. Unpublished training materials.

# Discussant's Comments

**Glynn Ligon**

## **Evaluation Software Publishing**

*Editors Note: The following is from a tape recording of the discussant's comments. It is intended to capture the prepared and spontaneous comments of a discussant to the material presented at the session. It was not written as a paper submitted for this volume.*

Wow! This has been an interesting session. The title went off on how test publishers make sure that their items are correct and not biased all the procedures are followed and then we ended up with No Child Left Behind. If we can focus on the test bias and item bias issues then we'll have more time to discuss the papers in more detail.

It's seems that what I hear is that bias is bad, but discrimination is good for tests and the test options. Because we wouldn't be giving tests if we weren't trying to discriminate in terms of performance or whatever we are measuring. So I think it is good, and probably not a coincidence, that the word discrimination of test items and tests was that people really talked about the test bias.

I have a quiz for you. In the sessions I've in there have been discussions about interesting analytical techniques. And so I have the names of four techniques, analysis techniques, that could be applied to this problem. And one of them is actually a

software product, and the other three are made up. You have to guess which one is which. Ready? First, we had a session that was talking about getting deeper into analysis, data mining and some things that might apply to test bias as well. And so, the first nominee for 'is this real' is a data mining tool called Polyanalyst. Then when you are on the discussion about how to query data, how to, try to get at what the meanings are within the databases. And that nominee is called Madame Query. Third, there is another one that came up later called the hierarchical linear supermodel. That's are third nominee, and the fourth nominee came from the Indiana section where the presented discussed the Indiana model or presenting data on their web. And the title of that analysis is the Indiananalysis 500. So now you have to guess which one of these is real. Polyanalysis, Madame Query, hierarchical linear supermodel, or Indianaanalysis 500?

Anybody want to take a chance? Hierarchical Linear Supermodel? No. Polyanalyst actually is real, that's the answer. Okay so there is bias in that because either you were in a session earlier and you heard it or you had some knowledge and actually got it right because you knew something or because you can see through all of the others. But I think it's interesting to think about item bias and when it's good and when it's not.

I think it's tremendous that No Child Left Behind has finally made item bias and test bias important because we're going to leave no child left behind and we have a goal of 100 percent of the students proficient. We can't afford to have bias in our items. We can't afford to have bias in tests, against any group of students. So I think No Child Left Behind has done a tremendous favor for our period and for the area of test and measurement.

One of the controversies around the stakes of working with No Child Left Behind has been determining which subgroups have enough in them to be statistically reliable to determine adequate yearly progress. Many of the states, in fact, the majority of the

states, have adopted a confidence interval which is based on sampling theory. Sampling theory says that the students at a given school at a given year were randomly selected from the general population of students and that's why they were in that school. Now here is the problem. Talk with any principal, and that principal is going to tell you that no way the students in his/her school were randomly selected from any population.

The problem with test bias is pretty much the same thing, and that is what is the population or the subgroups that we are trying to not be biased against. How do we define those? These subgroups are becoming so diverse themselves that is much harder now to really know who we are trying not to be biased against.

I really like what Steve said, when he pointed out that when he was talking about performance assessment and what we've learned from a decade of work. A lot of people tried their very best to make performance assessments work. And that the difference between the different groups of test takers turned out to be higher on performance assessments than they were on multiple choice tests. I've been an advocate of multiple choice tests for many years and it's really interesting on how no matter what the attack has been on multiple choice tests, they always kind of stay in there and we always rediscover why we have multiple choice tests. And in this case we have multiple choice tests because they are good at being scored. Machines are not very biased when they start scoring items whereas the scores on performance assessments is very problematic.

Steve and Marge, I've really enjoyed what you all described. I know that a tremendous effort was made in making sure that these tests and the items on the tests are as fair as possible. My own personal example was when my son, who is now in college, was a little squirt and he was taking an IQ test. I was on the other side of the wall listening and one of the items was 'look at these pictures and point to the skillet.' And our son just didn't know. First of all, we don't cook much, and so we didn't have

a skillet. And he'd never heard of the word skillet and so his IQ was lower because his parents didn't cook and we didn't use skillets in our house. Well, how do you control for that? We would assume that the test was not very biased against the middle class.

Then there was my favorite vocabulary test that we administered in Austin. If you've ever been in Austin, Texas and read the newspaper you know its called the Austin American Statesman. And the vocabulary item for these middle school students asked 'what is a statesman?' So how do you develop a test that is not going to discriminate against the students in Austin as opposed to Houston or another Texas city when you do a state graduation exam.

And then I think Steven said that we should consider including some diverse cultural content so that the students will feel some connection with the test. And that's one for the bar later because I don't see how you do that without creating more bias.

Jennifer, I think pointed out that when you started with at biased against groups you have to define a lot of groups – SES, LEP, student's IEP, etc. No Child Left Behind identifies gender, race and ethnicity. But where do you stop? Do you include kids who come from homes where they don't cook? Where do you stop identifying the groups that you don't want to be biased against? And that becomes more and more problematic.

There was a quick mention about different modes of testing. I think maybe that that changes the bias game a little, but many states bar some different modes such as the use of online tests. Is there going to be a difference in terms of which groups might be discriminated against, or for, with online testing?

Okay, so is test bias playing a role with adequate yearly progress? I came in hoping to hear an answer to that. Is there something that test bias or item bias that's playing a

role in the determination of adequate yearly progress? I think the bottom line is it has to be. Yes there probably is, but we don't really know if the states are looking at that very closely. But what the states have done is they have been very, very good at putting in their contracts to the test developers to make sure that all of these steps have been followed so that they will win in court if a court challenge comes. Most states have been extremely successful with this. So I think that when we look at test bias we have to give the test publishers and state education agencies a lot of credit for working hard enough that they can convince the federal courts and state courts that these tests aren't biased.

Many years ago I was in a school down on the border of Texas and Mexico. And I was sitting in a principal's office talking about a couple of students coming back and forth across the border every two weeks. He explained his problems with this family and moving around and not being stable in a school. Finally the principal said you know sometimes I think we can explain away so much of a kids' problems at school that we really don't stop and just help them. And I think that some of that relates to test bias as well. We have to be careful not to explain everything away, and not interpret the real differences that the different groups have on these tests.

So, I want to say that I think No Child Left behind is going to be great. And I don't agree with Bob Linn and with what Jennifer was saying earlier. The standard, really I think that is a misunderstanding about NCLB but let me go through a couple examples now that I've said that.

The 100 percent goal of the year 2014 is a mistake and I think that is going to be changed. This whole issue arose because people are sitting in this room as part of a committee that have set 100 percent goal for graduation or zero percent for dropout. That's when they are in those situations they say how can you say any child is going to fail and have a goal that is less than 100 percent. And so you set it so far out down the

road that you stretch everybody and their goals, but you know that in the future that the goal of 100 percent is going to change. I think that is a positive.

No Child Left Behind requires an adequate yearly progress destination for every single school in the nation even if they are too small to have enough students in the subgroups. So nobody is supposed to escape. Schools and districts with too few students to be part of this statistical calculation have to get some destination for adequate yearly progress. So, that's another misconception that a lot of people don't understand.

In the area of curriculum you have to get the kids to pass the tests. I did my professional growing up in Texas, one of the first states to get these tests and make them high stakes. And I watched schools that were doing an incredibly bad job turn around and start doing an incredibly good job with some very low performing students. To me, the problem which, I call educational malpractice, is when our educators take kids who already can pass the test and they are teaching them to prepare for the test. That's not the test's fault. Those kids should go on and do other things in their curriculum. There are so many students out there that can't pass the test yet that need that type of instruction. So we have to be more creative, and a little more logical how we are delivering the instruction. I personally don't think the curriculum is being dumbed down. The curriculum can get the students to pass the test. Now, we need to deliver it to every student, and if you are above that point you've still got the rest of the curriculum there that the teacher should be teaching.

But what I think is really great about No Child Left Behind, is that we're actually tackling the issues that we're all talking about here at AERA. There aren't many sessions that aren't No Child Left Behind related in some way, including this one on test bias. And I think that's great that a federal education law that's come around that's forced us to agree, disagree, and talk about all of these issues. And I think that a session like this is tremendous too, because it is a topic that doesn't need to be

forgotten, that does not need to be pushed aside. It does need to be at the front of the minds of the people in this room because we're the ones that need to watch for the test bias in our assessments. Besides our goal really shouldn't be just the win in court, the goal should be to win in the schools so students can walk across the stage at graduation.