Cultural Diversity and Ethnic Minority Psychology
2014, Vol. 20, No. 4, 570–582

1099-9809/14/$12.00 DOI: 10.1037/a0036503

Exploring the Various Interpretations of "Test Bias"

Russell T. Warne
Utah Valley University

Myeongsun Yoon
Texas A&M University

Chris J. Price
Utah Valley University

Test bias is a hotly debated topic in society, especially as it relates to diverse groups of examinees who often score low on standardized tests. However, the phrase "test bias" has a multitude of interpretations that many people are not aware of. In this article, we explain five different meanings of "test bias" and summarize the empirical and theoretical evidence related to each interpretation. The five meanings are as follows: (a) mean group differences, (b) differential predictive validity, (c) differential item functioning, (d) differing factor structures of tests, and (e) unequal consequences of test use for various groups. We explain in this article why meanings (a) and (e) are not actual forms of test bias and that there are serious concerns about (b). In our conclusion, we discuss the benefits of standardized testing for diverse examinees and urge readers to be careful and precise in their use of the phrase "test bias."

Keywords: test bias, differential item functioning, item bias, group differences, standardized tests

> "That's a great deal to make one word mean," Alice said in a thoughtful tone.
>
> "When I make a word do a lot of work like that," said Humpty Dumpty, "I always pay it extra." (Carroll, 1871/1917, p. 100)

In the English language there are many words and phrases that receive "extra pay" from those who—like Humpty Dumpty—give them many different meanings. One phrase in particular that has received a great deal of "extra pay" is *test bias*, which we found to have a multitude of meanings (Reynolds & Lowe, 2009). The purpose of this article is to explore the nature of test bias from the perspective of psychometrics (i.e., the science of mental testing). We intend to technically define the phrase and expound upon five different ways that *test bias* is often used, discuss the nature of item content as it relates to bias, and the benefits of standardized testing for diverse examinees—especially in the realm of education. We believe that the exploration of the phrase is important because test bias is a hotly debated topic in education and psychology, but some of these debates have not been productive because of those expressing opposing sides are often using the phrase differently (e.g., the exchange between Mercer, 1979, and Clarizio, 1979).

Although an article about semantics and terminology would itself be useful, it would probably be of limited interest to the readers of *Cultural Diversity and Ethnic Minority Psychology*. Therefore, we also discuss in this article the benefits of standardized educational and psychological tests for diverse examinees and those who advocate for diverse populations. We hope that our discussion of test bias will empower advocates of marginalized groups and improve the quality of discourse about the use of tests in psychology and education.

What Is a Standardized Test?

Before discussing the phrase *test bias*, it is imperative to define the phrase *standardized test*. In the popular lexicon the phrase refers to a government-mandated multiple-choice test in which the examinees (usually students) are required to give their responses by filling in a bubble on an answer sheet with a No. 2 pencil. This definition is, however, limiting. The word *standardized* merely refers to the fact that administration, format, and scoring of a test are the same for all examinees—which is an essential requirement to producing interpretable data (Sireci, 2005). There is nothing about this standardization that dictates that tests must be multiple choice format, or that a bubble sheet must be used, or that a standardized test be required or administered by a government entity. Indeed, standardized tests have many item formats, require examinees to make a wide variety of responses, and are administered by many persons and organizations.

Standardized tests have diverse formats. Some, such as licensure exams for architects or physicians, require examinees to create a product or perform a task under a uniform set of constraints. Others—such as the Torrance Test of Creative Thinking—require examinees to produce their own responses, which are then graded through a strict scoring rubric. Indeed, because consistency of format, scoring, and administration are the only requirements for a test to be standardized, one could argue that many life tasks are standardized. Despite the broad definition of the phrase *standardized test*, most authors who raise concerns about test bias are doing

Russell T. Warne, Department of Behavioral Science, Utah Valley University; Myeongsun Yoon, Department of Educational Psychology, Texas A&M University; Chris J. Price, Department of Behavioral Science, Utah Valley University.

Correspondence concerning this article should be addressed to Russell T. Warne, Department of Behavioral Science, Utah Valley University, 800 W. University Parkway, MC 115Orem, UT 84058. E-mail: rwarne@uvu.edu

so in the context of high-stakes standardized tests, such as college admissions tests, high school exit exams, intelligence tests, and employment tests.

Definitions of Test Bias

According to the standards set by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), *test bias* ". . . is said to arise when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 74). Although this definition is helpful, it is also quite broad and open to a variety of interpretations. In our experiences with the topic, we have found five ways that the phrase "test bias" has been interpreted in the literature:

1. Score gaps between groups which result from members of one group—on average—scoring higher than members of another group.
2. Differences in the ability of scores to predict outcomes for examinees.
3. Items functioning differently for examinees who belong to different groups.
4. Differences in the intercorrelations and groupings (i.e., factor structure) of items.
5. Consequences of test use or interpretation that create or perpetuate social inequalities between groups.

These methods of observing test bias are presented in the approximate order that these arguments first appeared in peer-reviewed journals because often later conceptualizations of "test bias" were often created to compensate for difficulties or shortcomings associated with earlier interpretations of "test bias." We also think that the order of interpretation of these definitions is pedagogically appropriate because (except for Interpretation 5) the explanations and examples of the "test bias" become increasingly more technical and complex as one moves down the list.

In this section of the article, we will examine each of these interpretations and explain mainstream thought among psychometricians concerning each and how they relate to AERA et al.'s, 1999 broad definition. It is important for readers to realize as they read this article that the groups referred to in the list above do not have to be racial groups, although racial and ethnic groups get the most attention in conversations about test bias. Other possible groups may be gender, religious, age, economic, education, cultural, national, or any other type of group imaginable.

Interpretation #1: Mean Score Differences

One of the most consistent—and frustrating—findings in quantitative educational studies is the pervasive score differences among racial and ethnic groups. On many academic achievement and aptitude tests, Asian American students score higher than White students, who then in turn score higher than Hispanics and African American students. These results have been found on intelligence tests (e.g., Gottfredson, 1997; Neisser et al., 1996; Roid, 2003), academic aptitude tests (e.g., Lohman, 2005), tests for identifying gifted children (e.g., Olszewski-Kubilius & Lee, 2011) and children with special needs (e.g., Morgan, Farkas, Hillemeier, & Maczuga, 2012), high school exit exams (e.g., Nichols, 2003), standardized academic achievement tests (e.g., Hoover et al., 2003; Forsyth et al., 2003; Lee, 2002), college admissions tests (e.g., Flowers, 2008; Posselt, Jaquette, Bielby, & Bastedo, 2012), and employment tests (O'Boyle & McDaniel, 2009). Moreover, these score gaps—especially those between African American and White examinees—are long-standing and have been observed in every generation since the beginning of modern mental testing in the United States (e.g., Cleary, Humphreys, Kendrick, & Wesman, 1975; Lee, 2002; Terman, 1928; Yoakum & Yerkes, 1920). Score gaps among racial and ethnic groups are so pervasive that their existence has been discussed at length in official APA publications (Cleary et al., 1975; Neisser et al., 1996) and has been called ". . . not a debatable issue" (Kaplan & Saccuzzo, 2009, p. 512; see Rushton & Jensen, 2005, p. 236, for almost identical language).

While the existence of score gaps on tests among demographic groups is well established, the *causes* of such group differences is still strongly debated (Reynolds, 2000). One common explanation for score differences on academic tests among demographic groups is test bias, with critics contending that deficiencies in the test cause the score gaps and make scores incomparable across demographic groups (as per AERA et al. 1999, definition of test bias). Richert, for example, wrote,

> "Measures of academic achievement that are most often used by schools [to identify gifted students], including teacher recommendations, grades, and especially standardized tests, have been amply demonstrated to have cultural biases . . ." (Richert, 2003, pp. 150–151, emphasis added).

According to Richert, not only are tests biased, but the evidence is overwhelming that bias is an inherent characteristic of standardized tests. Similarly, Salend and his colleagues stated, "Research indicates that *norm-referenced standardized tests are culturally and socially biased . . .*" (Salend, Garrick Duhaney, & Montgomery, 2002, p. 290, emphasis added). Beliefs about the inherently biased nature of tests are also found among other authors (e.g., Mensch & Mensch, 1991), including some that are highly respected in their fields (e.g., Ford, 2003; Gould, 1981). Such claims are common in the journalistic media, too (Cronbach, 1975; Gottfredson, 1994; Phelps, 2003; Reynolds, 2000).

Others have a much more sinister view of standardized testing. Moss described her experience teaching at a high school where, "Most of my students were poor and African American . . ." (p. 217). She stated,

> By the end of 13 years of experience, I became convinced that it did not matter how successful students of color became, the test would be revised to insure we start over in the cyclical process of teaching students how to demonstrate their ability to take culturally biased standardized tests. (Moss, 2008, p. 217)

For Moss, standardized tests are not just biased as some accident of their creation. Rather, the writers of the tests her students took were nefarious in their work, and the test creators intended to use the tests to discriminate against her students (see Carter & Goodwin, 1994; Mercer, 1979; and Smith, 2003; for a similar viewpoint of standardized tests).

What all of these examples have in common is that the writers believe that because the average scores for some groups is consistently below the average scores of other groups, the test must be biased and the scores from different demographic groups therefore have different meanings across groups. The group mean differences interpretation of *test bias* is probably the most popular interpretation of the five discussed in the article. This interpretation is illustrated in Figure 1, which shows two normal distributions. Although the two score distributions overlap, there is a difference of roughly half a standard deviation between the average scores in these groups. Because of this mean difference, there could be detrimental consequences to disadvantaged group members if the test is used to select students for an educational or therapeutic program.

Nevertheless, this interpretation of test bias is unanimously disregarded by those with training in psychological testing. Reynolds, for example, stated, "The mean differences definition of test bias is *by far* [emphasis added] the most widely rejected of all definitions of bias by psychometricians who study these issues" (Reynolds 2000, p. 146). Many other testing experts agree (e.g., AERA et al., 1999; Camara, 2009; Clarizio, 1979; Flaugher, 1978; Linn & Drasgow, 1987; Reynolds & Lowe, 2009).

Some advocates for diverse populations may be perplexed by this stance among psychometricians. If these tests consistently produce score gaps among groups, how could the tests *not* be biased? A simple thought experiment should answer this question. Instead of Figure 1 referring to two demographic groups' score distributions on an academic test, one should imagine that it refers to two groups' score distributions on a test of job satisfaction. The reader should then imagine that the lower distribution represents scores on a test of job satisfaction obtained from medical interns and the higher distribution should represent scores obtained from tenured college professors. A typical intern's schedule includes 80 hr of work per week, nights on call, and very stressful working conditions, while tenured university faculty have a great deal of work flexibility, high job security, and tend to enjoy their jobs. Under these circumstances, the professors *should* outscore the medical interns on a test of job satisfaction. Any other results would lead a reasonable observer to strongly question the validity of the test results. The lesson from this thought experiment is that mean score gaps are not evidence of test bias because there may be other explanations of score gaps. In fact, score gaps may indicate that the test is operating exactly as it should and measures real differences among groups—as is the case with this hypothetical test of job satisfaction (Clarizio, 1979; Linn & Drasgow, 1987).

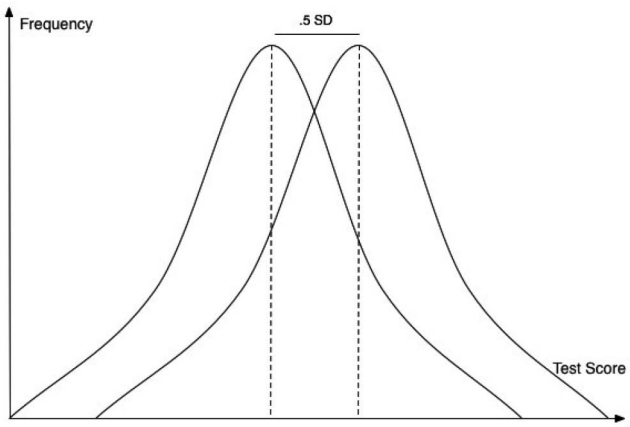

Figure 1. Score distributions for two groups. The mean difference between the two groups' scores is .5 *SD*.

Interpretation #2: Differential Predictive Validity

Many educational tests are used to predict an outcome, such as success in an educational program. If a test score is able to predict some future outcome (called a *criterion*), it is said to have *predictive validity*. For example, the SAT and ACT are designed to predict a student's probability of succeeding in college. Although details vary from study to study, both tests tend to be moderately strong predictors of college grades ($r \approx .40$), although the predictive power of these tests increases when combined with information about the student's high school grades (Camara, 2009; Kaplan & Saccuzzo, 2009; Maruyama, 2012; Zwick, 2006). College admissions test score correlations with other criteria are less consistent because other outcomes (e.g., income, career success, or whether a student will graduate with a bachelor's degree) occur many years after a student's score is obtained—which means many circumstances can intervene before the outcome is observed. Restriction of range of data has also been shown to reduce the strength of the correlations between test scores and outcomes (Camara, 2009; Zwick, 2006, 2007).

If mean score gaps themselves are not evidence of bias, perhaps a test could be biased if it is better at predicting outcomes for some groups and worse at predicting outcomes for other groups, a situation called *differential predictive validity* (Camilli, 2006). In other words, if the predictive validity of the test score varies from group to group, then it is an indication that it is not appropriate to use the test to make predictions for at least some examinees and scores may have different meanings across groups, making the test biased (AERA et al., 1999, p. 79). A few possible visual representations of differential predictive validity across groups are shown in Figure 2, a–c. Ideally, if a test has equal predictive validity for both groups, then the same regression line will apply to both groups, as demonstrated by Figure 2a. As is apparent in Figure 2a, one group scores higher than the other on average, but test scores predict outcomes for both groups equally well. In this case, no test bias would exist (Reynolds & Lowe, 2009).

However, if mean differences in scores exist between two groups, the schematic shown in Figure 2b is also possible. In the figure, there are two scatterplots. However, the two groups cannot be represented by the same regression line. Instead, there are two parallel regression lines that each separately represent the relationship between the test score and the outcome for each group. The dotted line halfway between the two main regression lines represents the regression line that would apply to both groups combined. As is apparent, the overall regression line represents the predicted outcomes for neither group very well. This would be an example of differential predictive validity (Cleary, 1968).

The difference in predictive validity shown in Figure 2b occurs somewhat frequently in academic testing, such as intelligence tests (Reschly & Sabers, 1979) and preschool assessment batteries (Reynolds, 1980). However, most testing experts are not terribly concerned when this differential predictive validity occurs because

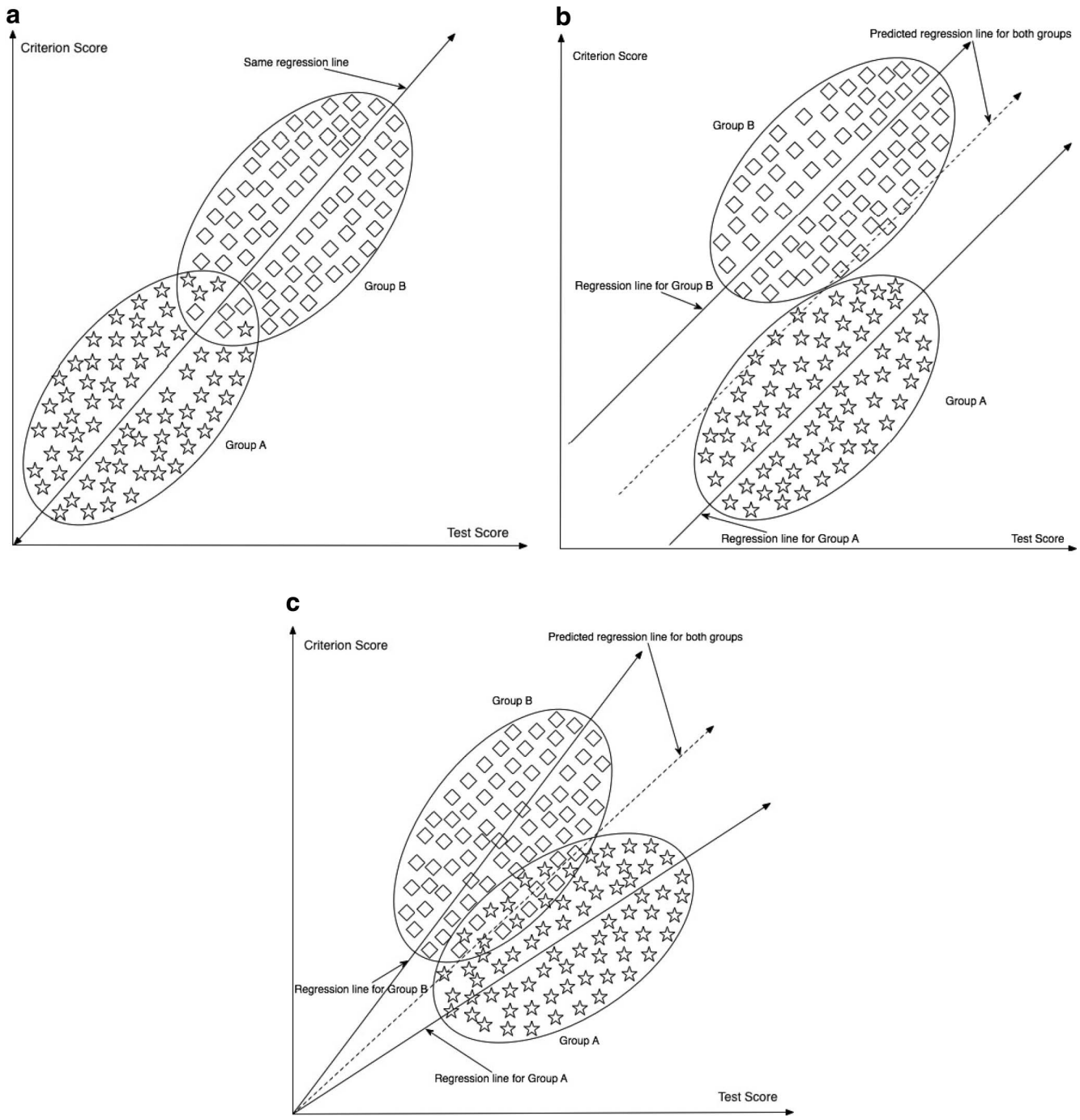

Figure 2. (a) Scatterplots for the scores of two groups that share the same regression line. Notice that Group A on average obtains lower scores than members of Group B. (b) Scatterplots for the scores of two groups with different, but parallel, regression lines. The middle dashed line represents the regression line for both groups combined. Notice that Group A on average obtains lower scores, but the combined regression line predicts more favorable outcomes than would be expected from a regression line solely based on Group A's data. (c) Scatterplots for the scores of two groups with different, nonparallel regression lines. The middle dashed line represents the regression line for both groups combined.

it does not hurt disadvantaged groups. This is because if outcomes are predicted using the regression equation from both groups combined, then the predictions for disadvantaged group members will systematically be *more favorable* than the predictions based on the disadvantaged group's regression data alone (Clarizio, 1979; Crocker & Algina, 2008; Kaplan & Saccuzzo, 2009). This can be seen in Figure 2b, which shows that for a given test score the predicted outcome for a disadvantaged examinee will always

be higher using the regression equation for the combined groups than for the regression equation based on just the disadvantaged group's data.

In the real world of educational testing, this is exhibited by data from the SAT. Young (2004) has shown that predictions of African American and Hispanic students' college grades when using data from the entire body of SAT examinees are systematically higher than diverse students' actual college grades. In other words, by using a test on which they score lower than other groups to decide who is admitted to college, African American and Hispanic students actually benefit and are admitted to college at greater rates than they would be if an admissions test were designed specifically for these diverse groups. This paradox has been long recognized by testing experts (e.g., Cleary, 1968; Cronbach, 1980).

Another reason differential prediction validity does not bother many experts is because some outcomes are influenced by a wide variety of independent variables (not just test scores), and these other variables may impact groups of examinees in different ways. For example, the SAT exhibits differential predictions for college graduation of White and Black examinees, with the latter being predicted to graduate at lower rates than White examinees, even if the students obtain the same SAT score (Zwick, 2006). This may be because of test problems, but it may also be because Black students face more barriers to graduation than White students. It has been suggested that a hostile campus environment, greater financial problems, greater anxieties about academic abilities, and other life challenges may make graduation more difficult for Black students compared with their White classmates who obtain the same score on the SAT (Zwick, 2007). Therefore, differential validity as displayed in Figure 2b—like mean score differences among groups—by itself is not evidence of test bias because it can have causes other than problems with the test.

Finally, there is some evidence that the type of differential predictive validity shown in Figure 2b is merely a statistical artifact. Traditional ordinary least squares (OLS) regression operates under the assumption that the predictor variable is measured perfectly (Young, Kane, Monfils, Li, & Ezzo, 2013). However, this is definitely not true with test scores, which always have some error (Kaplan & Saccuzzo, 2009) that can distort predictions made with OLS regression (Kane & Mroch, 2010). Indeed, in multiple studies (e.g., Kane & Mroch, 2010; Young et al., 2013) when researchers used an alternative prediction method that took into account test score error, the differential predictive validity of the SAT for two different racial groups disappeared. These results indicate that differential predictive validity may sometimes be a result of the statistical model that a researcher chooses—not an actual difference among examinee groups in test performance.

A more troubling example of differential predictive validity, however, is found in Figure 2c. Like Figure 2b, it shows differing regression lines for two groups of examinees. However, the regression lines in this example are radically different because they are not parallel. This leads to inaccurate predictions for all students when a single regression equation based on all examinees is used (like in Figure 2b). In this case, the test score does not make the same predictions for the two groups (Kaplan & Saccuzzo, 2009). Thus, combining their data into one general regression line is theoretically untenable. This form of differential prediction validity is insurmountable and has no other solution than to interpret group members' scores separately, although how to use those scores may be complicated and unclear (Allen & Yen, 1979). However, the differential predictive validity results shown in Figure 2c are rare and almost never encountered with real data from professionally developed tests (Kaplan & Saccuzzo, 2009) because the differences in correlations between test scores and criteria rarely differ across groups by more than what would be expected from regular sampling error (Jensen, 1998).

Interpretation #3: Differences in Group Performance on Specific Items

The first two interpretations of test bias discussed in this article—mean score gaps and differential predictive validity—are concerned with total test scores. However, tests are comprised of individual items. Therefore, the question of bias is perhaps best examined at the item level. Experts who subscribe to this belief have developed procedures to measure *differential item functioning* (DIF, also called *item bias*), which occurs when an item performs differently for different groups. For DIF to be present two examinees who belong to different groups but *with equal levels of individual ability* must have different probabilities of correctly answering an item (AERA et al., 1999; Camilli, 2006; Cleary & Hilton, 1968; McDonald, 1999; Swaminathan & Rogers, 1990). Items may display DIF for a variety of reasons. For example, an item

> . . . may be measuring something different from the remainder of the test or it may be measuring with different levels of precision for different subgroups of examinees. Such an item may offer a valid measurement of some narrow element of the intended construct, or it may tap some construct-irrelevant component that advantages or disadvantages members of one group. (AERA et al., 1999, pp. 77, 78)

Procedures for examining DIF are very complex and technical. Put in simple terms, they usually split examinees into groups based on their total test score. Then for each item, the probability that members of each group of interest (e.g., racial groups, gender groups, cultural groups, socioeconomic groups) will correctly answer an item correct is calculated. If the null hypothesis that the two groups have an equal probability of answering the item correctly is rejected (i.e., $p < .05$, .01, or some other predetermined value for α), then DIF is said to be present and the test scores are biased in the AERA et al. (1999) sense of the phrase in that the scores do not have the same meaning for both groups. For example, if average-scoring low-socioeconomic status (SES) and high-SES examinees have the same probability of answering an item correctly, then DIF is not present for that particular item. This summary of DIF suffices for the purposes of this article, but readers should recognize that DIF analysis takes many forms. The reader is encouraged to examine more thorough and technical treatments of DIF, some of which do not rely on null hypothesis testing (e.g., Camilli, 2006; Crocker & Algina, 2008, chapter 16; Embretson & Reise, 2000, pp. 249–263; Reynolds & Lowe, 2009, pp. 345–351; Swaminathan & Rogers, 1990).

DIF is quite useful for helping researchers and test developers find items that do not function the same way across demographic or social groups. Past research has shown, for example, that DIF items that favor diverse students occur when the items contain content that is of interest or especially relevant to members of their demographic group—such as reading passages about the civil

rights movement for Black students (Schmitt & Dorans, 1990; Zwick, 2007). Similarly, Hispanic students tend to find verbal analogy items that use words that have Spanish cognates (i.e., words in Spanish that are spelled similarly to English words with the same meaning) easier than White students (Zwick, 2006, 2007). For math items, basic computation items tend to favor Black students, while White students perform better on story problems (Hunter & Schmidt, 2000; Zwick, 2007).

DIF across gender groups has also been shown in certain types of items on academic tests. Zwick (2007, p. 26) explained,

> Women tend not to do as well as a matched group of men on verbal SAT items . . . about scientific topics or about stereotypically male interests, like sports or military activities . . . On the other hand, women tend to perform better than their male counterparts on questions about human relationships or questions about the arts . . .

However, these tendencies are the exception to the rule; usually it is not clear why an item displays DIF (AERA et al., 1999; Camilli, 2006; Flaugher, 1978; Reynolds & Lowe, 2009; Schmitt & Dorans, 1990; Zwick, 2007).

If an item displays DIF, then it may be eliminated from a test, revised, or be balanced with DIF that cancels out the DIF in the first item (Hunter & Schmidt, 2000; Zwick, 2006; see Hoover et al., 2003 for an example of these procedures being used on an academic achievement test). If many items are found to exhibit DIF that consistently favors one group over another, then the entire test score may systematically favor one group solely because of group membership (and not the relevant characteristic being tested, such as academic achievement). Such results would be a strong indication that the test cannot be interpreted in the same way for both groups.

How common is DIF? With the plethora of tests in the world, most of which consist of many items, it should be unsurprising that many tests have at least a few items that display DIF. However, in professionally developed tests—such as the high-stakes educational tests that are the subject of conversations about test bias—items with large DIF are not present. This is because professional test standards of practice and ethics to which test developers are bound require that items with substantial DIF be removed from tests (AERA et al., 1999). It is common, however, for items with modest DIF to be found on professional tests. For example, on Form A of the Iowa Tests of Educational Development there are 1,134 items across all grade levels. Of these, 5 (0.4%) favor females, 13 (1.1%) favor males, 3 (0.3%) favor Black students, 1 favors White students over Black students (0.1%), none favor Hispanic students, and 2 (0.2%) favor White students over Hispanic students (Forsyth et al., 2003, p. 86). As is apparent in this example, items displaying DIF on professionally created tests are usually so few and the DIF so small that the impact on total scores is negligible (Camara, 2009; Goodman & Hambleton, 2005; Reynolds, 2000; Reynolds & Lowe, 2009).

DIF procedures are probably the most common way that tests are examined for bias today. However, the procedure is not without its problems. DIF is inherently atheoretical and subjected to statistical results that incorrectly flag acceptable items as having DIF (Hunter & Schmidt, 2000). It is also fundamentally circular; in DIF procedures based on total test scores, examinees must be matched on their total score in order to examine a particular item for bias. But the total score usually includes the item being screened for DIF. Therefore, one must accept that the total test score is unbiased as a prerequisite to test individual items for bias; then, once DIF is not found in any of the items, a researcher can assert that the total test score is not biased. This circular reasoning (i.e., where the test score must be assumed to be unbiased in order to determine that items are unbiased) has been correctly criticized by researchers (e.g., Navas-Ara & Gómez-Benito, 2002).

Many DIF procedures also create problems statistically; if one item with DIF is found but the total test score used to match examinees across groups, then the advantage that a group receives from the DIF item must be balanced out by other items that favor the other group(s). This creates statistical artifacts in which items are labeled as incorrectly having DIF (Andrich & Hagquist, 2012). These are issues that the testing field is still grappling with, and fully satisfactory solutions have not yet been found. DIF procedures based on latent variable methods such as confirmatory factor analysis and item response theory are promising and have reduced the severity of these statistical and logical problems. Nevertheless, DIF has shed light on important questions about test construction and has undoubtedly made tests fairer today than they were in previous generations.

Interpretation #4: Differing Factor Structures

Test items are rarely interpreted in isolation, but rather items are combined together in order to produce a score that is subject to interpretation. In order for the interpretation or scoring of a set of items to make sense, the items must intercorrelate with one another to form a coherent group; these groups of items are assumed to be related to one another because similar responses are caused by an underlying *factor*, such as "verbal ability," "agreeableness," or "mathematics achievement." Most tests are created with a theory concerning the number of factors and which items belong to each factor (Kaplan & Saccuzzo, 2009). This theory of the makeup of the factors is essential for interpretation because if items do not intercorrelate according to theory, then the test score(s) may not make sense and may be uninterpretable.

Factors are estimated and interpreted by researchers using a multivariate statistics method called *factor analysis* (Gorsuch, 2003). The results of a factor analysis show the number of factors in a set of test items and which items belong to each factor. Figure 3a shows an example of seven items (represented by the seven rectangles) that form two factors (represented by the two circles). As is apparent in the figure, one factor consists of four items and the other factor consists of three items. The double-headed arrow between the two circles shows that the two factors are correlated.

Factor analysis is relevant to the argument of test bias because it is possible that the items may form different factors for different groups (Reynolds & Lowe, 2009). This is tested through a method called a *test of factorial invariance*, which is a complex statistical process of testing whether the test items have the same factor structure for both groups (Meredith, 1993; Millsap & Everson, 1993; Reise, Widaman, & Pugh, 1993; Schmitt & Kuljanin, 2008). If items on a test have the same factor structure for both groups, then no bias is present (Reynolds & Lowe, 2009).

However, if—as a hypothetical example—test items for one group have the factor structure in Figure 3a and the same items have the factor structure in Figure 3b in another group, then test bias. This would signify that the items from the test do not

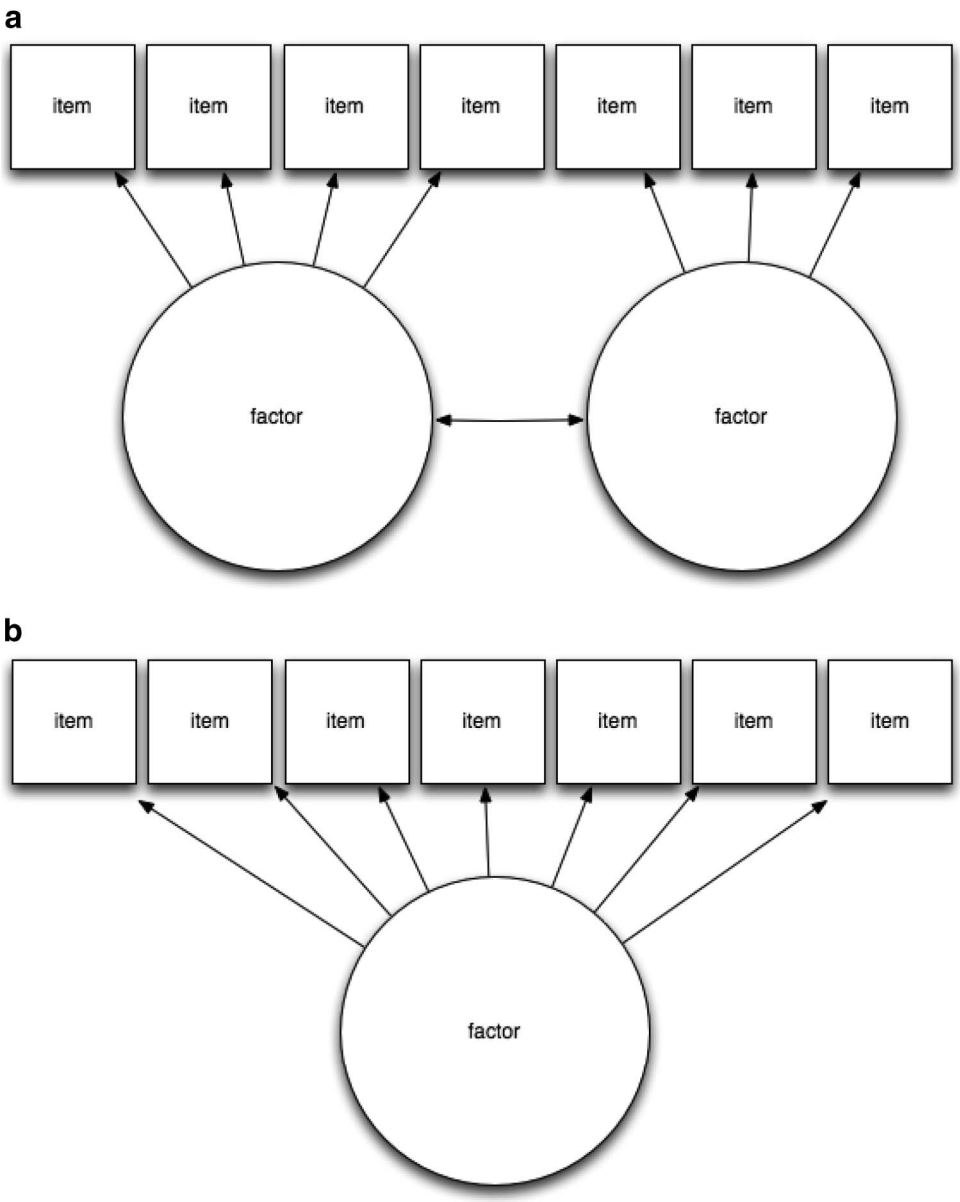

Figure 3. (a) A 7-item test with two factors: four items load onto the first factor and three items load onto the second factor. (b) A 7-item test with all items loading on a single factor.

necessarily "behave" the same way for both groups of subjects and test bias—as defined by AERA et al. (1999)—may be present because the scores from the two groups would have different meanings. Differing factor structures may indicate a number of possibilities, including the following:

- The test items may be interpreted differently by the two different groups.
- The psychological construct (e.g., depression, personality, intelligence, language arts achievement) may have different structures for the two groups. The nature of the construct may vary across groups because of cultural, developmental, or other differences.
- The test may measure completely different constructs for the two groups.
- Groups may use different mental processes to respond to the items.
- Examinees in different groups use different strategies when responding to test questions.

If two different groups of examinees have different factor structures for a group of items, it is not appropriate to interpret the items in the same way for both groups and scores cannot be compared across groups (Meredith, 1993; Reynolds & Lowe, 2009), a situation that meets AERA et al.'s (1999) definition of test bias. Examples abound of tests that have differing factor structure items across demographic groups, especially for instruments created by nonpsychometricians (e.g., Li et al., 2009; Warne, 2011). For most professionally developed intelligence, academic, and aptitude tests, tests of invariance usually indicate that the factor structure is

the same across groups (e.g., Beaujean, McGlaughlin, & Margiules, 2009; Benson, Hulac, & Kranzler, 2010; Dolan, 2000).

Evaluations of factor structure are especially complex and require large datasets (Meredith, 1993). However, they are generally agreed to be the best methods of evaluating test bias (Borsboom, 2006). Because tests of invariance are somewhat new, there are still several aspects of them that remain unresolved. First, the results of a test of invariance are frequently not as clear and unambiguous as the hypothetical example shown in Figure 3, a and b. Rather, differences among factor structure are often a matter of degree, not of kind. It is often difficult to know what to do with a test when some parts of it operate the same across demographic groups and other parts do not (Millsap & Kwok, 2004). Also—just like DIF—it is not always clear why factor structures vary across tests or why parts of a test function differently for different groups (Schmitt & Kuljanin, 2008).

Interpretation #5: Unequal Consequences of Test Use Across Groups

Others claim that a test is biased if its use disadvantages some groups in society compared with others. Such a test is said to have poor *consequential validity* because the consequences of using the test are socially undesirable. For example, women on average score consistently higher on the Beck Depression Inventory (BDI) than men, indicating higher levels of depression (Santor, Ramsay, & Zuroff, 1994). Therefore, women may be more likely to be diagnosed as depressed and be administered therapy and prescription drugs. Some may believe that the negative consequences for women of using the BDI to diagnose depression are undesirable and damaging to American society because they may foster sexism, stigma, and discrimination in individual examinees' lives. These people say that the consequences of using the BDI are too great and that the test is biased because it has poor consequential validity. The consequential validity perspective of test bias is unique because it is a subjective judgment based on ethical and moral values, whereas the other interpretations of test bias are based on—and can be evaluated with—statistical data.

Those who make the argument that standardized tests that produce negative consequences for some segments of society are making the argument that using the tests is *unfair*. Because of the moral nature of fairness, the question of whether it is fair to use a standardized test—even a high-quality, professionally developed standardized test that exhibits no DIF or differential predictive validity and has the same factor structure for all examinees—cannot be settled scientifically. Rather, such judgments are best made by society through the public mechanisms that exist in democratic nations to make decisions where competing values among people lead to tensions. These mechanisms include state and federal legislatures, the ballot box, the court system, and the public meetings of elected and appointed officials in which administrative rules are created. Expert analysis of these public discussions (e.g., Buckendahl & Hunt, 2005; Kaplan & Saccuzzo, 2009; Phelps, 2003; Phillips, 2000; Phillips & Camara, 2006; Schafer, 2000; Ward, 2000) often leads to fascinating understandings of society's current and past values as they relate to standardized testing—especially in employment and education.

We believe that it is easy to make a very compelling case that it is unfair to use a test to gather information to determine whether a person goes to college, is diagnosed with a mental illness, gets hired for a job, graduates from high school, and so forth. However, we—along with many other testing experts (e.g., AERA et al., 1999; Camilli, 2006; Kaplan & Saccuzzo, 2009)—strongly believe that *test bias* and *fairness* are not synonymous. Mainstream professional opinion about this topic is that test bias is a purely technical term—defined by Interpretations 2–4—concerned with actual test score interpretation (Popham, 1997), which is the crux of the AERA et al. (1999) definition of test bias that we subscribe to. Therefore, moral or ethical judgments concerning the unfairness of test use are not evidence that the test is biased, although the judgments may be used to construct persuasive arguments that a test is unfair. This important distinction leads to the widely recognized (among psychometricians) fact that an absence of test bias is a prerequisite for fairness (AERA et al., 1999; Kaplan & Saccuzzo, 2009). However, it is possible for a test to demonstrate no statistical evidence of bias, yet still be unfair in the eyes of some members of society.

But Look at the Items! It's Obvious That They're Biased!

Some critics claim that many standardized tests are biased use test items themselves as the basis for their arguments against tests. These people claim that the test items ask questions about cultural information that diverse examinees cannot relate to and/or are much less likely to be exposed to. For example, in the civil rights case *Larry P. v. Wilson Riles* (1979) the plaintiffs' lawyers—who were advocating for African American students in California public schools—examined the revised version of the Wechsler Intelligence Scale for Children (WISC–R) and found items that they claimed were biased against African American schoolchildren. According to Elliott (1987), one of these items was, "Who wrote *Romeo and Juliet?*" The lawyers argued that such cultural information was irrelevant to diverse examinees who may have had fewer opportunities than White students to learn the answer to this test question. Therefore, it was discriminatory to use the item to test students' intelligence. In the technical language of testing, it would be said that this item would lack *face validity* for testing the intelligence of diverse children. Such face validity arguments are made solely on the basis of whether an item appears to examine what test creators purport it examines (Carlson & Geisinger, 2009; Gottfredson, 2009; Kane, 2006).

Such an argument about test bias is so self-evident to critics of standardized tests that merely the existence of such items is enough evidence that the tests are biased (Elliott, 1987). However, this is a problematic stance to take for multiple reasons. First, such items are somewhat rare on tests, and like the number of items displaying DIF, are not numerous enough to explain the score differences that exist among demographic groups on most standardized tests (Flaugher, 1978). Second, such arguments against standardized tests based solely on face validity judgments are "anecdote wholly unsupported by evidence" (Elliott, 1987, p. 123). Clarizio (1979), agreed stating, "Subjective, armchair analysis of bias—the primary method used by critics who charge cultural bias—is no substitute for item statistics" (p. 81; see also Mercer, 1979). We contend that critics who make arguments about test bias based on single items forget Aristotle's maxim that, "One swallow does not a summer make" (as quoted in Johansen, 1998, p. 382).

In other words, a single incident—in this case one unfair item—is not indicative of a trend of a biased test (Carlson & Geisinger, 2009; Reschly, 1980).

Third, it is notoriously difficult to pick out items that function differently across groups through face validity judgments. In one well known attempt at trying to "eyeball" biased items, Judge John Grady in the *PASE v. Hannon* case examined every item on the WISC–R and Stanford-Binet Intelligence Scale and found seven items on the WISC–R to be biased and one item on the Stanford-Binet to exhibit bias (Bersoff, 1981). In later DIF analyses, however, it was found that *none* of the items that Judge Grady had identified as being biased actually exhibited DIF (Koh, Abbatiello, & McLoughlin, 1984). This demonstrates a well-known principle of psychometrics: subjective face validity judgments are prone to erroneous conclusions (Reschly, 1980) and that, "Face validity . . . does not offer evidence to support conclusions drawn from test scores" (Kaplan & Saccuzzo, 2009, p. 136).

Although subjective judgments about bias on the basis of face validity are inadequate for identifying test bias, examinations of experts are still an important part of the process to make tests as fair as possible for all examinees. Current ethical standards for testing demand that, "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 . . ." (AERA et al., 1999, p. 82). The most typical method of eliminating such offensiveness is through a *sensitivity panel*, which is an independent group of trained reviewers who examine every component of a test for language that might be offensive, threatening, controversial, stereotypical, sexist, insensitive, condescending, inflammatory, or distracting to any subgroup of examinees (Camilli, 2006; Reynolds, 2000; Reynolds & Lowe, 2009). Sensitivity panels also look for language or vocabulary that would have a different meaning for any subgroup (AERA et al., 1999) and therefore change the meaning of the item for some examinees. Sensitivity panels are usually made up of experts who are often themselves members of diverse racial, ethnic, gender, religious, or disability groups, and are carefully trained in identifying problematic language that would disadvantage an examinee.

Although not able to identify items that show DIF (Flaugher, 1978; Reynolds, 2000), sensitivity panels are an important aspect of test development that ensure that tests are as fair as possible for many groups of examinees. This is because sensitivity panels ensure that examinees are not exposed to distracting language—an important function in our opinion because we believe that every examinee has a right to be tested free from unnecessary distraction. We also think that any test that contains insensitive language is unfair, regardless of what statistical analyses may say about the items. Thankfully, the prevailing ethical standards of the testing field agree with us (AERA et al., 1999), and it is probably impossible today for a commercial testing company to sell an educational or employment test today without all components of a test being screened by a sensitivity panel (for examples of documentation of the work of sensitivity panels, please see ACT, 2007; Forsyth et al., 2003; and Hoover et al., 2003).

Benefits for Diverse Populations of Standardized Testing

Now that we have established various interpretations of the phrase *test bias* and discussed the strengths and weaknesses of each, we believe that it is important to discuss the benefits of standardized testing for diverse examinees because of the tendency of some advocates for diverse populations to either dismiss standardized tests as inherently biased (e.g., Ford, 2003; Richert, 2003; Smith, 2003) or as tools of powerful racial and economic groups to maintain their dominance over diverse populations (e.g., Carter & Goodwin, 1994; Mercer, 1979; Moss, 2008). We reject these positions—as do other testing experts (e.g., Gottfredson, 2009; Reynolds, 2000)—and believe that standardized tests can be quite useful in promoting the efforts of diverse populations in psychological, employment, and educational testing for two reasons: (a) standardized tests can be used as a measurement of social equality, and (b) standardized tests are less problematic than other alternatives. Each of these issues will be discussed briefly in the succeeding subsections.

Standardized Tests Can Measure Social Inequality

Many experts—including ourselves—are greatly concerned about the inequalities that exist in education and economic outcomes among different groups in society. We also believe that education can be a means of reducing inequalities among racial, gender, and other demographic groups. Standardized tests can be a yardstick to help advocates understand the degree of inequality among groups and how much progress society has made toward reducing these inequalities (Flaugher, 1978). As recognized by the authors of the current ethical standards for educational and psychological testing, "Properly designed and used, tests can and do further societal goals of fairness and equality of opportunity" (AERA et al., 1999, p. 73).

An example of the use of standardized tests as a measurement of inequality is the National Assessment of Educational Progress (NAEP), a testing program administered by the federal government. NAEP is important to educational researchers and policymakers because it is the only standardized test that produces results which permit comparisons across demographic groups and state lines (Lane et al., 2009). Moreover, federal law makes NAEP a sanctioned instrument for measuring educational progress (Koretz, 2003). Recent NAEP results displayed in Figure 4 show that since 1971 reading achievement has increased for White, Hispanic, and Black students. As encouraging as these results are, a more encouraging fact is that in that time the achievement gap between White students and the other two groups on NAEP reading assessments has dramatically narrowed (National Center for Educational Statistics, 2009). Similar results have also been found on NAEP mathematics scores since 1973 (National Center for Educational Statistics, 2009). Although progress has been made, it is clear from NAEP data that substantial score gaps in achievement still exist and that equality in achievement has not yet been obtained. This is information that would not be available without standardized tests.

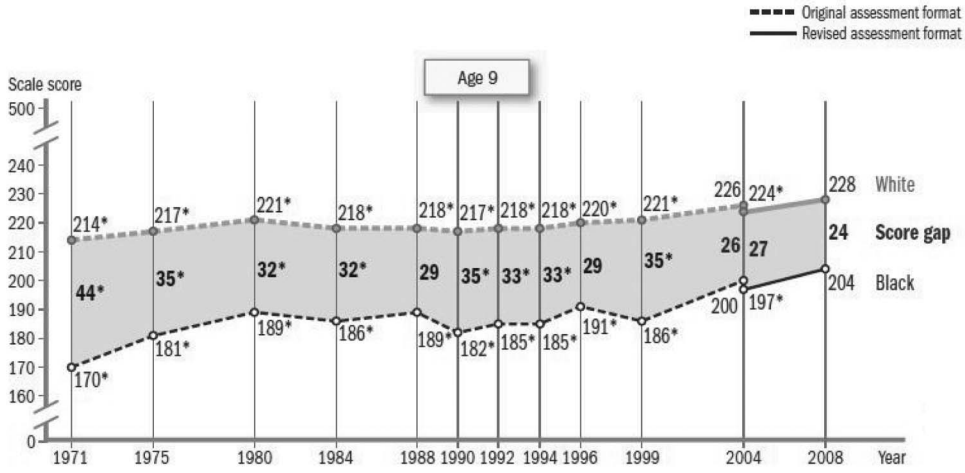

Figure 4. Trend in White–Black National Assessment of Educational Progress (NAEP) reading average scores and score gaps for 9-year-old students. Source: National Center for Education Statistics (2009, p. 14).

Standardized Tests Are Less Problematic Than Alternatives

Some critics of standardized tests wish to replace the tests with alternatives (such as diagnostic interviews, teacher ratings, grade point averages, and portfolio assessments) that they claim would be less discriminatory against diverse examinees. The quality of these alternatives is often greatly overstated and replacing standardized tests with them would cause far more problems than they would solve (Clarizio, 1979; Flaugher, 1978; O'Boyle & McDaniel, 2009; Gottfredson, 2009; Phelps, 2003). For example, some claim that using high school grade point averages (GPAs) alone in college admissions would be fairer than SAT scores. However, research has shown that using high school GPA as a predictor for college outcomes (such as GPA) results in greater prediction errors than using SAT scores as a predictor (Mattern, Shaw, & Kobrin, 2011; Zwick, 2007). Moreover, the recent trends in grade inflation (e.g., Camara, 2009; Posselt et al., 2012) and the strong tendency for teachers to be subjective and idiosyncratic in their awarding of grades (e.g., Cizek, Fitzgerald, & Rachor, 1995) have been well documented. These problems—and others that have been explained at length elsewhere (e.g., Camara, 2009; Zwick, 2006, 2007)—are much less severe with standardized tests than with any other alternative that has yet been suggested (Camara, 2009; Clarizio, 1979; Phelps, 2003; Reynolds & Lowe, 2009). Stated quite simply, standardized tests are the most efficient, cheapest, and least biased way of assessing a large number of people. This has been true since the 1920s.

Conclusion

We now end this article the same way we began it—with an excerpt from *Through the Looking-Glass*:

> "When *I* use a word," Humpty Dumpty said, in rather scornful tone, "it means just what I choose it to mean—neither more nor less."
>
> "The question is," said Alice, "whether you *can* make words mean so many different things." (Carroll, 1871/1917, p. 99)

Our intention in this article is not to play the role of Humpty Dumpty and arbitrarily choose what *test bias* does and does not mean. Instead, we hope in this article to clarify the meaning of a phrase that often receives "extra pay" because it means "a great deal" (Carroll, 1871/1917, p. 100). As defined by AERA et al., (1999), test bias exists when test scores have different meanings for different groups of examinees. As we have demonstrated, there are at least five common interpretations for the phrase *test bias*: (1) mean score differences between groups, (2) differential predictive validity, (3) differential item functioning, (4) differences in item factor structure, and (5) consequences of a test that disadvantage members of some demographic groups. This multitude of interpretations introduces a degree of impenetrability to conversations on the topic of test bias which we hope to clarify.

During the course of this article we have shown that some of these meanings of "test bias" are not supported: interpretation (1) is not supported by psychometricians because sometimes group scores differences are expected, while interpretation (5) is not concerned with test score meaning, so some experts do not consider it a manifestation of test bias. However, interpretations (2), (3), and (4) are potential manifestations of test bias (as defined by AERA et al., 1999) because they are possible ways that differences in score meaning can be detected statistically. We hope that we have explained the merits and drawbacks of each one and have added to the conversation about testing and will foster a more productive conversation about test use in society.

We encourage readers to use the information in this article for two purposes: first, to be more careful in their use of the phrase *test bias*. Although *test bias* does have each of the five meanings we have elucidated, haphazard use of the phrase introduces confusion. Careless or vague use of language is a sign of sloppy thinking (Woodford, 1967), which can only muddle dialogues among professionals about tests and their uses. We encourage readers to be precise with their language and to perhaps use terms like "differential predictive validity," "differential item functioning," "mean group differences," and so forth, in leiu of the term "test bias." Second, we hope that readers of this article will use the explanations we have given to use tests to advocate for diverse groups in society in a thoughtful and productive way. We believe that such advocacy efforts would be far more productive than attacking standardized tests.

References

ACT. (2007). *The ACT technical manual*. Retrieved from http://www.act.org/aap/pdf/ACT_Technical_Manual.pdf

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

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

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics, 37,* 387–416. doi:10.3102/1076998611411913

Beaujean, A. A., McGlaughlin, S. M., & Margulies, A. S. (2009). Factorial validity of the Reynolds Intellectual Assessment Scales for referred students. *Psychology in the Schools, 46,* 932–950. doi:10.1002/pits.20435

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment, 22,* 121–130. doi:10.1037/a0017767

Bersoff, D. N. (1981). Testing and the law. *American Psychologist, 36,* 1047–1056. doi:10.1037/0003-066X.36.10.1047

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440. doi:10.1007/s11336-006-1447-6

Buckendahl, C. W., & Hunt, R. (2005). Whose rules? The relation between the "rules" and "law" of testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 147–158). Mahwah, NJ: Erlbaum.

Camara, W. J. (2009). College admission testing: Myths and realities in an age of admissions hype. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 147–180). Washington, DC: American Psychological Association. doi:10.1037/11861-004

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: Praeger.

Carlson, J. F., & Geisinger, K. F. (2009). Psychological diagnostic testing: Addressing challenges in clinical applications of testing. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 67–88). Washington, DC: American Psychological Association. doi:10.1037/11861-002

Carroll, L. (1871/1917). *Through the looking-glass and what Alice found there*. New York, NY: Rand McNally.

Carter, R. T., & Goodwin, A. L. (1994). Racial identity and education. *Review of Research in Education, 20,* 291–336. doi:10.2307/1167387

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3,* 159–179. doi:10.1207/s15326977ea0302_3

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

Cleary, T. A. (1968). Test bias: Prediction of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124. doi:10.1111/j.1745-3984.1968.tb00613.x

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28,* 61–75. doi:10.1177/001316446802800106

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30,* 15–41. doi:10.1037/0003-066X.30.1.15

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30,* 1–14. doi:10.1037/0003-066X.30.1.1

Cronbach, L. J. (1980). Selection theory for a political world. *Public Personnel Management, 9,* 37–50.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35,* 21–50. doi:10.1207/S15327906MBR3501_2

Elliott, R. (1987). *Litigating intelligence: IQ tests, special education, and social science in the courtroom*. Dover, MA: Auburn House.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Flaugher, R. L. (1978). The many definitions of test bias. *American Psychologist, 33,* 671–679. doi:10.1037/0003-066X.33.7.671

Flowers, L. A. (2008). Racial differences in the impact of participating in Advanced Placement programs on educational and labor market outcomes. *Educational Foundations, 22,* 121–132.

Ford, D. Y. (2003). Equity and excellence: Culturally diverse students in gifted education. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (pp. 506–520). Boston, MA: Allyn & Bacon.

Forsyth, R. A., Ansley, T. N., Feldt, L. S., & Alnot, S. D. (2003). *Iowa Tests of Educational Development guide to research and development*. Itasca, IL: Riverside Publishing.

Goodman, D., & Hambleton, R. K. (2005). Some misconceptions about large-scale educational assessments. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 91–110). Mahwah, NJ: Erlbaum.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 143–164). Hoboken, NJ: Wiley.

Gottfredson, L. S. (1994). Egalitarian fiction and collective fraud. *Society, 31,* 53–59. doi:10.1007/BF02693231

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. *Intelligence, 24,* 79–132. doi:10.1016/S0160-2896(97)90014-3

Gottfredson, L. S. (2009). Logical fallacies used to dismiss the evidence on intelligence testing. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 11–65). Washington, DC: American Psychological Association. doi:10.1037/11861-001

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

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., . . . Shannon, G. P. (2003). *Iowa Tests of Basic Skills guide to research and development*. Itasca, IL: Riverside Publishing.

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6,* 151–158. doi:10.1037/1076-8971.6.1.151

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

Johansen, K. F. (1998). *A history of ancient philosophy from the beginnings to Augustine*. New York, NY: Routledge.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kane, M. T., & Mroch, A. A. (2010). Modeling group differences in OLS and orthogonal regression: Implications for differential validity studies. *Applied Measurement in Education, 23,* 215–241. doi:10.1080/08957347.2010.485990

Kaplan, R. M., & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth.

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

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice, 22,* 18–26. doi:10.1111/j.1745-3992.2003.tb00124.x

Lane, S., Zumbo, B. D., Abedi, J., Benson, J., Dossey, J., Elliott, S. N., . . . Willhoft, J. (2009). Prologue: An introduction to the evaluation of NAEP. *Applied Measurement in Education, 22,* 309–316. doi:10.1080/08957340903221436

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31,* 3–12. doi:10.3102/0013189X031001003

Li, H., Lee, D., Pfeiffer, S. I., Kamata, A., Kumtepe, A. T., & Rosado, J. (2009). Measurement invariance of the Gifted Rating Scales–School

Form across five cultural groups. *School Psychology Quarterly, 24,* 186–198. doi:10.1037/a0017382

Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and Practice, 6,* 13–17. doi:10.1111/j.1745-3992.1987.tb00405.x

Lohman, D. F. (2005). Review of Naglieri and Ford (2003): Does the Naglieri Nonverbal Ability Test identify equal proportions of high-scoring White, Black, and Hispanic students? *Gifted Child Quarterly, 49,* 19–28. doi:10.1177/001698620504900103

Larry P. v. Riles, 495 F. Supp. 926 (N. D. Cal. 1979).

Maruyama, G. (2012). Assessing college readiness: Should we be satisfied with ACT or other threshold scores? *Educational Researcher, 41,* 252–261. doi:10.3102/0013189X12455095

Mattern, K. D., Shaw, E. J., & Kobrin, J. L. (2011). An alternative presentation of incremental validity: Discrepant SAT and HSGPA performance. *Educational and Psychological Measurement, 71,* 638–662. doi:10.1177/0013164410383563

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Mensch, E., & Mensch, H. (1991). *The IQ mythology: Class, race, gender, and inequality*. Carbondale, IL: Southern Illinois University Press.

Mercer, J. R. (1979). In defense of racially and culturally nondiscriminatory assessment. *School Psychology Review, 8,* 89–115.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543. doi:10.1007/BF02294825

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334. doi:10.1177/014662169301700401

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9,* 93–115. doi:10.1037/1082-989X.9.1.93

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2012). Are minority children disproportionately represented in early intervention and early childhood special education? *Educational Researcher, 41,* 339–351. doi:10.3102/0013189X12459678

Moss, G. (2008). Diversity study circles in teacher education practice: An experiential learning project. *Teaching and Teacher Education, 24,* 216–224. doi:10.1016/j.tate.2006.10.010

National Center for Educational Statistics. (2009). *NAEP 2008 trends in academic progress* (NCES Report 2009–479). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2008/2009479.pdf

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment, 18,* 9–15. doi:10.1027//1015-5759.18.1.9

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101. doi:10.1037/0003-066X.51.2.77

Nichols, J. D. (2003). Prediction indicators for students failing the State of Indiana high school graduation exam. *Preventing School Failure, 47,* 112–120. doi:10.1080/10459880309604439

O'Boyle, E. H., Jr., & McDaniel, M. A. (2009). Criticisms of employment testing: A commentary. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 181–197). Washington, DC: American Psychological Association. doi:10.1037/11861-005

Olszewski-Kubilius, P., & Lee, S.-Y. (2011). Gender and other group differences in performance on off-level tests: Changes in the 21st century. *Gifted Child Quarterly, 55,* 54–73. doi:10.1177/0016986210382574

Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, NJ: Transaction Publishers.

Phillips, S. E. (2000). *GI Forum v. Texas Education Agency*: Psychometric evidence. *Applied Measurement in Education, 13,* 343–385. doi:10.1207/S15324818AME1304_04

Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733–755). Westport, CT: Praeger.

Popham, W. J. (1997). Consequential validity: Right concern–wrong concept. *Educational Measurement: Issues and Practice, 16,* 9–13. doi:10.1111/j.1745-3992.1997.tb00586.x

Posselt, J. R., Jaquette, O., Bielby, R., & Bastedo, M. N. (2012). Access without equity: Longitudinal analyses of institutional stratification by race and ethnicity, 1972–2004. *American Educational Research Journal, 49,* 1074–1111. doi:10.3102/0002831212439456

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566. doi:10.1037/0033-2909.114.3.552

Reschly, D. J. (1980). Psychological evidence in the *Larry P.* opinion: A case of right problem–wrong solution? *School Psychology Review, 9,* 123–135.

Reschly, D. J., & Sabers, D. L. (1979). Analysis of test bias in four groups with the regression definition. *Journal of Educational Measurement, 16,* 1–9. doi:10.1111/j.1745-3984.1979.tb00080.x

Reynolds, C. R. (1980). An examination of bias in a preschool battery across race and sex. *Journal of Educational Measurement, 17,* 137–146. doi:10.1111/j.1745-3984.1980.tb00822.x

Reynolds, C. R. (2000). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law, 6,* 144–150. doi:10.1037/1076-8971.6.1.144

Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 332–374). New York, NY: Wiley.

Richert, E. S. (2003). Excellence with justice in identification and programming. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 146–158). Boston, MA: Allyn & Bacon.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, fifth edition, technical manual*. Itasca, IL: Riverside Publishing.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11,* 235–294. doi:10.1037/1076-8971.11.2.235

Salend, S. J., Garrick Duhaney, L. M., & Montgomery, W. (2002). A comprehensive approach to identifying and addressing issues of disproportionate representation. *Remedial and Special Education, 23,* 289–299. doi:10.1177/07419325020230050401

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6,* 255–270. doi:10.1037/1040-3590.6.3.255

Schafer, W. D. (2000). *GI Forum v. Texas Education Agency*: Observations for states. *Applied Measurement in Education, 13,* 411–418. doi:10.1207/S15324818AME1304_07

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27,* 67–81. doi:10.1111/j.1745-3984.1990.tb00735.x

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18,* 210–222. doi:10.1016/j.hrmr.2008.03.003

Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 111–121). Mahwah, NJ: Erlbaum.

Smith, R. A. (2003). Race, poverty, & special education: Apprenticeships for prison work. *Poverty & Race, 12,* 1–4.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Terman, L. M. (1928). The influence of nature and nurture upon intelligence scores: An evaluation of the evidence in Part I of the 1928 Yearbook of the National Society for the Study of Education. *Journal of Educational Psychology, 19,* 362–373. doi:10.1037/h0071466

Ward, C. A. (2000). *GI Forum v. Texas Education Agency*: Implications for state assessment programs. *Applied Measurement in Education, 13,* 419–426. doi:10.1207/S15324818AME1304_08_1

Warne, R. T. (2011). An investigation of measurement invariance across genders on the Overexcitability Questionnaire–Two. *Journal of Advanced Academics, 22,* 578–593. doi:10.1177/1932202X11414821

Woodford, F. P. (1967). Sounder thinking through clearer writing. *Science, 156,* 743–745. doi:10.1126/science.156.3776.743

Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests*. New York, NY: Henry Holt and Company. doi:10.1037/11054-000

Young, J. W., Kane, M., Monfils, L., Li, C., & Ezzo, C. (2013, May). *Investigating differential validity and differential prediction using orthogonal regression analysis*. Paper presented at annual meeting of the American Educational Research Association, San Francisco, CA.

Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 289–302). New York, NY: Rouledge-Falmer. doi:10.4324/9780203463932_Differential_Validity_and_Prediction

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647–679). Westport, CT: Praeger.

Zwick, R. (2007). *College admission testing*. Retrieved from http://www.nacacnet.org/research/PublicationsResources/Marketplace/Documents/TestingWhitePaper.pdf