PERSONNEL PSYCHOLOGY
2007, 60, 165–199

UNDERSTANDING THE IMPACT OF TEST VALIDITY AND BIAS ON SELECTION ERRORS AND ADVERSE IMPACT IN HUMAN RESOURCE SELECTION

HERMAN AGUINIS
The Business School
University of Colorado at Denver and Health Sciences Center

MARLENE A. SMITH
The Business School
University of Colorado at Denver and Health Sciences Center

We propose an integrative framework for understanding the relationship among 4 closely related issues in human resource (HR) selection: test validity, test bias, selection errors, and adverse impact. One byproduct of our integrative approach is the concept of a previously undocumented source of selection errors we call *bias-based selection errors* (i.e., errors that arise from using a biased test as if it were unbiased). Our integrative framework provides researchers and practitioners with a unique tool that generates numerical answers to questions such as the following: What are the anticipated consequences for bias-based selection errors of various degrees of test validity and test bias? What are the anticipated consequences for adverse impact of various degrees of test validity and test bias? From a theory point of view, our framework provides a more complete picture of the selection process by integrating 4 key concepts that have not been examined simultaneously thus far. From a practical point of view, our framework provides test developers, employers, and policy makers a broader perspective and new insights regarding practical consequences associated with various selection systems that vary on their degree of validity and bias. We present a computer program available online to perform all needed calculations.

Human resource selection tests that are not supported by validity evidence are not useful in predicting job performance and other meaningful criteria. Tests that are biased are a legal liability and, in addition, using them can lead to unethical decision making. Consequently, test validity

We thank Rich Arvey, Phil Bobko, Peter Bryant, Larry James, Jim Outtz, and Frank Schmidt for their constructive comments on earlier drafts, and Kevin Turner for assistance with Java programming. An abbreviated version of this manuscript was presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX, May 2006.

Correspondence and requests for reprints should be addressed to Herman Aguinis, The Business School, University of Colorado at Denver and Health Sciences Center, Campus Box 165, P. O. Box 173364, Denver, Co. 80217-3364; http://www.cudenver.edu/∼haguinis.

and test bias are two of the most central concepts in human resource selection research and practice (e.g., Reilly & Chao, 1982; Schmidt, Pearlman, & Hunter, 1980). Although validity evidence can take on several forms, test validity is usually operationalized using a correlation coefficient (i.e., validity coefficient; Schmidt & Hunter, 1998). Similarly, although several definitions of test bias have been proposed (Darlington, 1971; Hunter & Schmidt, 1976; Petersen & Novick, 1976; Thorndike, 1971), potential test bias is usually assessed using a multiple regression framework in which race, sex, and other categorical variables related to protected class status are entered as moderators (AERA, APA, & NCME, 1999, Standard 7.6; Campbell, 1996; Cleary, 1968; Hough, Oswald, & Ployhart, 2001).

In addition to considerations regarding test validity and test bias, tests are most useful when they allow for selection decisions that minimize selection errors and avoid adverse impact. Selection errors occur when people who are hired do not meet performance standards (i.e., false positives) or when people are not hired but could have met performance expectations (i.e., false negatives; Cascio & Aguinis, 2005a, Chapter 13). Adverse impact is usually operationalized as a ratio of two selection ratios (SRs; Biddle, 2005; Bobko & Roth, 2004). Thus adverse impact is SR_1/SR_2, where SR_1 and SR_2 are the number of applicants selected divided by the total number of applicants for the minority and majority groups of applicants, respectively. It is desirable that adverse impact be as close to 1.0 as possible (e.g., for sex, similar selection ratios for men and women).

In spite of the voluminous literature on the related issues of test validity, test bias, selection errors, and adverse impact, researchers tackle these topics in isolation or in pairs. For example, researchers have studied the relationship between test validity and test bias (e.g., Darlington, 1971; Thorndike, 1971) and the relationship between test validity and selection errors (e.g., Curtis & Alf, 1969; Murphy & Shiarella, 1997). However, we have not been able to locate any published source that investigated the interrelationship among all four of these concepts explicitly. Moreover, some of the most widely read and cited books on personnel selection, staffing, and industrial psychology do not consider these concepts in an integrated manner. Instead, they typically discuss the concept of adverse impact in the chapter on legal issues, the topic of test bias in the chapter on fairness, and the topics of validity and selection errors in the chapter on prediction/decision making (e.g., Cascio & Aguinis, 2005a; Gatewood & Feild, 2001; Guion, 1998; Ployhart, Schneider, & Schmitt, 2006).

Human resource selection researchers and practitioners alike are clearly interested in the key and interrelated concepts of test validity, test bias, selection errors, and adverse impact. So, why is it that these four key concepts, although closely linked to each other, have been studied mainly in isolation or in pairs only? We believe this void in the literature is due to the absence of an integrative framework that would allow for an understanding of how these concepts are intrinsically and interactively related to each other. Such an integrative framework would provide a useful decision-making tool through which selection instruments could be evaluated before they are actually used to make decisions based on psychometric issues around the prediction of applicants' job performance as well as value-based considerations at the team, organizational, and societal levels associated with anticipated adverse impact. These value-based considerations can include achieving a balanced and diverse workforce and enhancing perceptions of justice among job applicants (Zedeck & Goldstein, 2000).

Accordingly, the goal of the present article is to propose an integrative framework that uses well-known regression and correlation principles and references to a standard normal table of probabilities for measuring the interrelationships among test validity, test bias, selection errors, and adverse impact. In the process of developing the framework, we discuss an often-unrecognized source of selection errors. Most human resource researchers and practitioners are familiar with selection errors that result from imperfect regression predictions, such as those described by Taylor and Russell (1939). In our framework, selection errors can also occur when biased tests are used as if they were unbiased. In this article, we refer to the former as *predictive selection errors* and the latter as *bias-based selection errors*. Our discussion of bias-based selection errors is particularly noteworthy given that a conclusion of no or small bias may be due to several methodological and statistical artifacts that reduce sample-based effect sizes substantially in relation to their population counterparts (Aguinis, Beaty, Boik, & Pierce, 2005) and also often lead to low statistical power (Aguinis & Stone-Romero, 1997). In particular, due to the numerous methodological and statistical artifacts that affect test bias assessment, it is possible that a test thought to be unbiased may actually be biased (Aguinis, 1995, 2004; Aguinis & Stone-Romero, 1997). Our analysis provides new insights into both positive and negative outcomes associated with the use of a selection tool that, unknown to the decision maker, is actually biased.

We advance a framework that integrates all four concepts within a single model, provides human resource selection researchers and practitioners with a new tool to look at key issues in human resource selection, and generates answers to questions such as the following, What are the consequences for bias-based selection errors of various degrees of test validity and test bias? What are the consequences for adverse impact of various degrees of test validity and test bias? Note that our integrative framework is needed because, although in some cases a researcher may be able to use data from an empirical validity study to compute adverse impact and predictive selection errors, sample sizes may not be large enough to obtain meaningful estimates of certain proportions (e.g., proportion of individuals whose scores fall above a cutoff score on a test but whose performance scores fall below a desirable level). Furthermore, empirical validity studies do not consider the issue of bias-based selection errors as incorporated in our framework. In short, our integration of key human resource selection concepts allows us to ask and answer questions that thus far were not possible.

From a theory point of view, our integrative framework provides a more complete picture of the selection process by integrating four key concepts that have not been examined simultaneously thus far. This integration will allow for fruitful areas of research in the future such as the development of selection tools that maximize validity, minimize bias, and mitigate adverse impact and selection errors. In addition, our proposed framework will allow researchers to better understand potential tradeoffs between test validity and test bias in affecting adverse impact. From a practical point of view, our framework provides test developers, employers, and the legal system a broader perspective regarding practical consequences associated with various selection systems that vary regarding their degree of validity and bias. We also present a computer program that can be executed online to implement our framework and perform all needed calculations. This online calculator can be used to anticipate how the numerical values of these key concepts change interactively before a selection test is actually used. Thus, the program can be used to evaluate the tradeoffs involved in maximizing job performance based on psychometric principles versus maximizing the influence of other important value-based principles associated with adverse impact (workforce integration and diversity, perceptions of justice of the selection system, etc.).

The article is organized as follows. First, we provide a description of our integrative framework, including definitions of its key components. We do so by minimizing the technical material (which is mostly presented in Appendixes A through C) and, instead, we present our framework using graphs. Second, we provide an analytic description of how to use our framework to derive precise numerical values for anticipated adverse impact and bias-based selection errors. Third, we describe three distinct selection scenarios to show the applicability of our framework to a diverse set of selection situations. We close by discussing the implications of using our proposed framework for theory, practice, and policy making. The final section of the paper also describes a computer program available online that produces graphs and performs all needed calculations.

Basic Concepts and Definitions: Test Bias, Expected Selection Ratios, Expected Adverse Impact, and Bias-Based Expected Selection Errors

Consider a situation in which applicants can be classified as belonging to one of two groups based on protected status (e.g., race or sex). In our presentation, Group 1 represents the minority group (e.g., ethnic minority) and Group 2 the majority group (e.g., ethnic majority). In some situations, Group 1 and Group 2 follow the same regression line, such as the one labeled common regression line: $E(Y\,|\,X) = \alpha + \beta X$ in Figure 1, which links test scores (X) and some criterion such as job performance (Y). (The other two regression lines in Figure 1 will be discussed shortly.) This common regression line represents an unbiased test because, at any given test score (e.g., x^* in Figure 1), it predicts identical performance levels (y^*) for both groups (AERA, APA, & NCME, 1999). Because an unbiased test is one in which both groups follow the same regression line, we refer to that line as the *common regression line*. We adopt the consensual operationalization of *test bias* as differences in regression lines across groups given that "Cleary's (1968) regression model of test bias or fairness has received the greatest acceptance and use among psychometricians" (Martocchio & Whitener, 1990, p. 489; see Aguinis, 2004; Campbell, 1996; Hough et al., 2001; and Maxwell & Arvey, 1993, for similar statements).

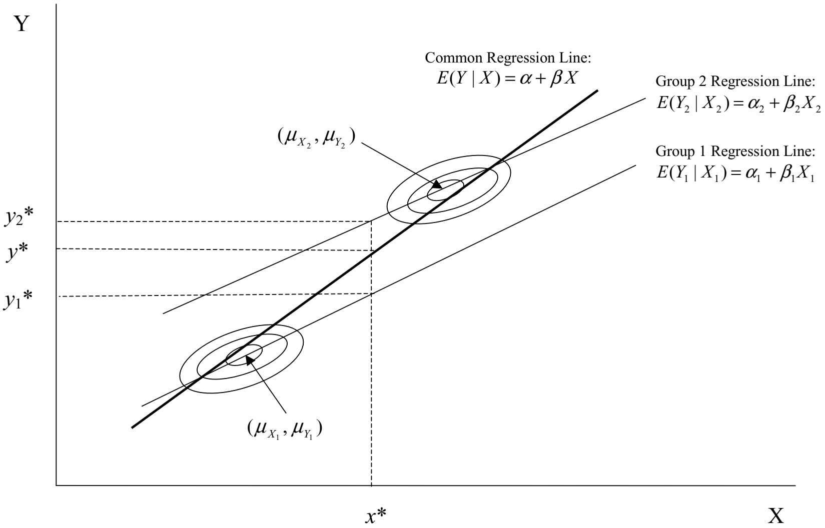


Figure 1: **Graphic Illustration of Expected Performance for Common and Group-Based Regression Lines.**

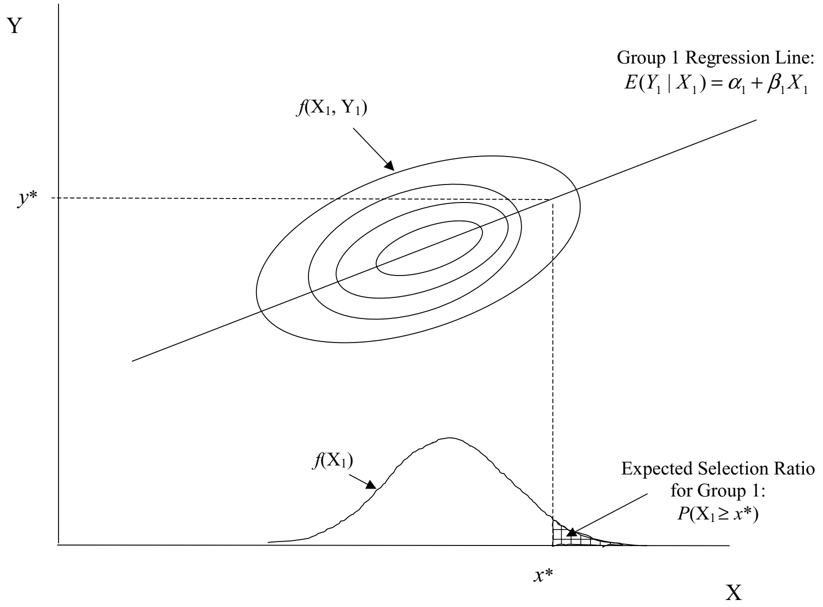


Figure 2: **Graphic Illustration of the Expected Selection Ratio for Group 1 (i.e., Ethnic Minority Group).**

Figure 1 also depicts a situation involving a biased test in which each group follows its distinct regression relationship, lines that we refer to as *group-based regression lines*. If a test is biased, it will predict average performance $y_1^* = E(Y_1 \mid x^*)$ for Group 1 and $y_2^* = E(Y_2 \mid x^*)$ for Group 2. The group-based regression lines in Figure 1 depict a fairly common finding regarding the use of cognitive ability tests in human resource selection: differences between groups are detected regarding intercepts (but not slopes) for the group-based regression lines (Hunter & Schmidt, 1976; Reilly, 1973; Rotundo & Sackett, 1999).

We begin our presentation with unbiased tests and by initially referencing Group 1 (i.e., the focal group, which is typically the minority group; Biddle, 2005). As shown in Figure 2, test and performance scores for Group 1 are presumed to follow a continuous bivariate distribution function. This function is labeled $f(X_1, Y_1)$, where the "1" subscripts reference Group 1. In practice, the decision maker may have a minimum desired performance level in mind, symbolized by y^* in Figure 2, and would like to offer employment to individuals whose desired performance is y^* or higher. At y^*, inverse prediction using the regression of Y_1 on

X_1 associates the desired performance level y^* with the selection cutoff x^* (Cascio & Aguinis, 2005b). Of course, some individuals whose test scores are x^* will perform better than, and some worse than, y^* because the validity coefficient is less than 1.0 in absolute value, and the regression model does not offer a perfect prediction mechanism. Similarly, some individuals whose test scores are lower than x^* will be able to perform at level y^* or higher. Therefore, in order to distinguish between individuals and averages, we define x^* in Figure 2 to be the *expected selection cutoff* given y^*. The expected selection cutoff is the organization's best guess as to the appropriate selection cutoff while in the planning stage of implementing selection decisions. If $E(Y_1 \mid X_1) = \alpha_1 + \beta_1 X_1$, at the specific value of y^*, $y^* = \alpha_1 + \beta_1 x^*$. Thus, the expected selection cutoff is found by solving for x^*:

$$x^* = (y^* - \alpha_1)/\beta_1. \tag{1}$$

We define the *expected hiring pool* to be the group of individuals with test scores of at least x^*. Referring again to Figure 2, the expected hiring pool includes all individuals whose test scores exceed x^* on the joint distribution, $f(X_1, Y_1)$. We define the *expected selection ratio* to be the area under $f(X_1, Y_1)$ to the right of x^* (i.e., the percentage of the population under consideration for employment). As shown in Equation A5 in Appendix A, this area is identical to the area to the right of x^* under the *marginal* distribution for Group 1's test scores, $f(X_1)$, an area we label $P(X_1 \geq x^*)$. An analogous definition applies to Group 2. Thus, a group's expected selection ratio is the upper tail area of its marginal distribution function of test scores at the expected selection cutoff. Our definition of the expected selection ratio is analogous to the more commonly understood definition of a selection ratio as the observed percentage of those hired relative to the total number of applicants. A key difference is that our framework allows decision makers to consider a priori (i.e., expected and *before* decisions are made) percentages as opposed to a posteriori (i.e., observed and *after the fact*) percentages.

We define *expected adverse impact* to be the ratio of the expected selection ratios for Groups 1 and 2. Figure 3 provides a graphic illustration of how to calculate expected adverse impact when a test is unbiased: expected adverse impact is simply the ratio of the smaller shaded area, $P(X_1 \geq x^*)$, to the larger shaded area, $P(X_2 \geq x^*)$. Thus, expected adverse impact (EAI) for an unbiased test is calculated as the ratio of two tail probabilities:

$$\text{EAI} = P\big(X_1 \geq x^*\big)/P\big(X_2 \geq x^*\big). \tag{2}$$

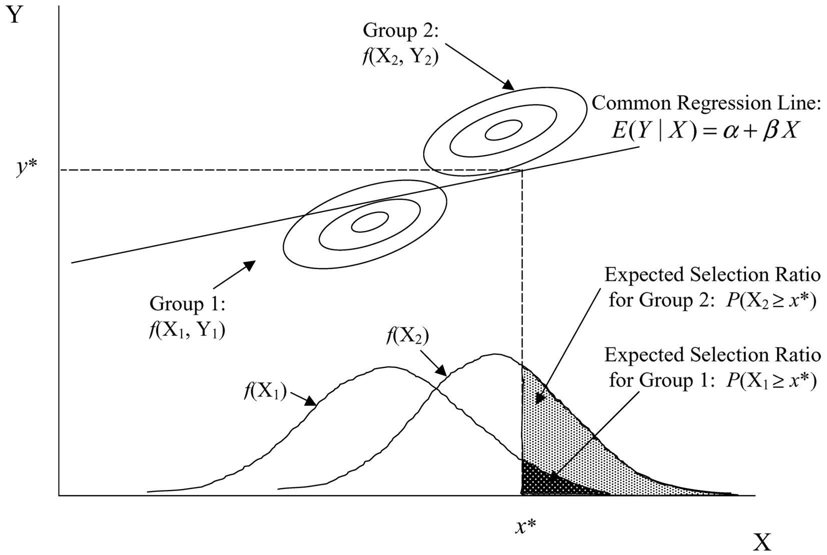


Figure 3: **Graphic Illustration of Expected Adverse Impact for an Unbiased Test.**

Now, consider the situation in which a decision maker mistakenly believes that he or she is using an unbiased test when in fact the test is biased. As noted earlier, this may not be an uncommon situation given the low statistical power of the test bias assessment procedures (e.g., Aguinis, 1995, 2004; Aguinis & Stone-Romero, 1997). In this case, the selection system will produce unanticipated selection errors (i.e., bias-based errors)—that is, expected false positives and expected false negatives that result from using a biased test as if it were unbiased. Biased-based selection errors are distinct from those arising from a test with less than perfect validity. To illustrate, Figure 4 shows a biased test in which the Group 1 regression line is different from that of Group 2. Superimposed on Figure 4 is the common regression line. In Figure 4, as in Figures 2 and 3, the desired performance level is y^*. Suppose that, believing that the test is unbiased, decision makers use the common regression line to choose x^* as the expected selection cutoff so that the expected hiring pool includes those individuals whose test scores equal or exceed x^*. What happens if, unknown to the decision makers, the test is actually biased? If the test is biased, there are two group-based regression lines, not one. Using the group-based regression lines instead of the common line in Figure 4 indicates that the true expected selection cutoff associated with y^* for Group 1 is not x^* but rather x_1^*. Therefore, if the common line is used, those individuals from Group 1 whose test scores fall within the range [x^*, x_1^*] are in the expected hiring pool but are not expected to reach performance level y^* because their performance, as predicted by the (true) Group 1 regression line, is less than y^*. We define such applicants as (bias-based) *expected false positives.* In this case, expected false positives occur because the true expected hiring pool at y^* is smaller than anticipated. The *probability of expected false positives* is the area under $f(X_1)$ between x^* and x_1^* (as shown in Figure 4). The precise numerical value for the probability of expected false positives is given by:

$$P\left(x^* \leq X_1 \leq x_1^*\right). \tag{3}$$

Refer again to Figure 4 and suppose that applicants from Group 2 whose test scores meet or exceed x^* are anticipated to be able to satisfy the minimum performance standard y^* based on the inverse prediction from the common regression line. If the true X–Y relationship for Group 2 is its distinct regression line (instead of the common regression line), then Group 2 applicants whose test scores fall within the range [x_2^*, x^*] are also expected to be able to meet the minimum performance level y^* but are not in the expected hiring pool. We define a (bias-based) *expected false negative* as an applicant who will not be considered for employment based on the common regression line but who is expected to be able to meet or exceed the minimum performance standard if the group-based line is used. In Figure 4, the *probability of expected false negatives* is the area under $f(\mathrm{X}_2)$ between x_2^* and x^*; that is

$$P\left(x_2^* \leq X_2 \leq x^*\right). \tag{4}$$

Given the situation in Figure 4, both groups will display expected false negatives whenever x_1^* and x_2^* are both less than x^* at y^*. Analogously, both groups will display expected false positives when x_1^* and x_2^* exceed x^* at y^*.

In general, the human resource selection literature refers to false positives and false negatives when selection predictions differ from actual selection outcomes due to the fact that no prediction system is perfect (i.e., validity coefficients are always less than 1.0 in absolute value). In this article, we provide a detailed analysis of an additional source of false positives and false negatives: selection errors that occur when a biased test is unknowingly used as if it were unbiased (i.e., bias-based errors). Thus, in the remainder of this article, we use the terms expected false positives and false negatives to refer to bias-biased selection errors.

Our integrative framework also leads to the conclusion that when biased tests are used as if they were unbiased, average performance levels by group will deviate, sometimes quite drastically, from desired performance

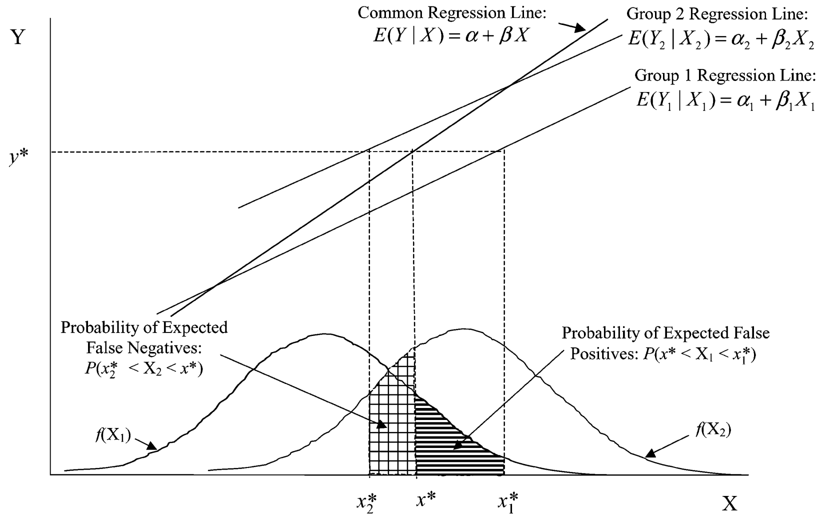

Y
y^*
Common Regression Line: $E(Y \mid X) = \alpha + \beta X$
Group 2 Regression Line: $E(Y_2 \mid X_2) = \alpha_2 + \beta_2 X_2$
Group 1 Regression Line: $E(Y_1 \mid X_1) = \alpha_1 + \beta_1 X_1$
Probability of Expected False Negatives: $P(x_2^* < X_2 < x^*)$
Probability of Expected False Positives: $P(x^* < X_1 < x_1^*)$
$f(X_1)$
$f(X_2)$
x_2^*
x^*
x_1^*
X

Figure 4: **Graphic Illustration of the Probabilities of Expected False Positives and Expected False Negatives.**

levels. Let's assume that selection decisions are made based on the common regression line when in fact the lines are not precisely identical across groups. Referring back to Figure 1, at x^*, decision makers expect performance level y^* for both groups because they believe that the test is unbiased (i.e., that the common regression line holds true). If instead there are distinct group-based lines, the actual average performance is y_1^* for Group 1 and y_2^* for Group 2. The further away are y_1^* and y_2^* from y^*, the greater will be the deviation of anticipated performance using the common regression line from predictions using the group-based lines. Using Figure 1 to illustrate, if x^* exceeds the point of intersection of the common regression line and the Group 2 regression line, then using the common regression line when the test is actually biased will lead to the unpleasant surprise that the performance of those people hired from both groups is not, on average, as good as anticipated. On the other hand, if x^* is less than the intersection of the common regression line and the Group 1 regression line, when the common line is used to make hiring decisions, decision makers will face the pleasant surprise that average observed performance of both groups will exceed anticipated performance.

Test validity (i.e., the correlation between X and Y) is part of our integrative framework via the shape of the ellipses for $f(X_1, Y_1)$ and $f(X_2, Y_2)$ in Figures 1 through 3. For example, as the validity coefficient for Group 1 approaches zero, its elliptical contours become more circular in shape. Thus, in our framework, test validity is used as an input to the model. Its value can change from situation to situation so that different values for the validity coefficient generate numerically different expected selection ratios, expected selection errors, and expected adverse impact.

Finally, we emphasize that the discussion above refers to *expected* selection ratios, *expected* selection cutoffs, *expected* adverse impact, and probabilities of *expected* false positives and negatives. Thus, our analysis allows researchers and practitioners to make decisions regarding the use of specific selection tools *before* actual outcomes are observed. This is obviously a key advantage of our framework in that it can be used in the planning stages of selection decision making and, thus, allows decision makers to be proactive and anticipate selection outcomes, some of them highly undesirable (e.g., unacceptable rates of expected false positives and negatives, severe expected adverse impact), before they are actually observed.

Putting the Basic Concepts Together: Obtaining Numerical Values Using the Normal Model

Once assumptions are made about the stochastic properties of $f(X_1, Y_1)$ and $f(X_2, Y_2)$, we can obtain specific numerical values for regression lines, expected selection ratios, expected adverse impact, and probabilities of expected false positives and negatives. In this section of the article, we presume that both bivariate distribution functions, $f(X_1, Y_1)$ and $f(X_2, Y_2)$, are normally distributed (which is a usual assumption in the human resource selection literature; e.g., Guion, 1998; Hunter, Schmidt, & Judiesch, 1990; Taylor & Russell, 1939; Thomas, 1990), with mean test scores μ_{X_1} and μ_{X_2}, mean performance scores μ_{X_1} and μ_{X_2}, test score standard deviations σ_{X_1} and σ_{X_2}, performance standard deviations σ_{Y_1} and σ_{Y_2}, and test validities ρ_1 and ρ_2. Regression lines for predicting Y from X for each group can be derived from these parameters as follows (see Appendix B for details):

$$\text{Group 1:} \quad E(Y_1 \mid X_1) = \alpha_1 + \beta_1 X_1 \tag{5}$$

where

$$\beta_1 = \rho_1(\sigma_{Y_1}/\sigma_{X_1}), \quad \text{and} \tag{6}$$

$$\alpha_1 = \mu_{Y_1} - \beta_1 \mu_{X_1}, \tag{7}$$

and similarly for Group 2. A test is unbiased if Groups 1 and 2 have identical regression lines, or $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. To determine whether a test is unbiased, we compare the numerical values of the intercepts α_1 and α_2 and of the slopes β_1 and β_2.

Suppose that the test is unbiased. Because we have invoked normality, the expected selection ratio for Group 1 (ESR_1) is found by referring to a standard normal table:

$$\mathrm{ESR}_{\mathrm{c}1} = P\left(X_1 \geq x^*\right) = P\left(Z \geq z^*_{c1}\right), \tag{8}$$

where the subscript "c" reminds us that the expected selection cutoff is found from the common regression line and

$$z^*_{c1} = \frac{x^* - \mu_{X_1}}{\sigma_{X_1}}. \tag{9}$$

For Group 2,

$$\mathrm{ESR}_{c2} = P\left(X_2 \geq x^*\right) = P\left(Z \geq z^*_{c2}\right), \tag{10}$$

where

$$z^*_{c2} = \frac{x^* - \mu_{X_2}}{\sigma_{X_2}}. \tag{11}$$

Because expected adverse impact (EAI) is the ratio of the expected selection ratios, it follows from Equations 8 and 10 that

$$\mathrm{EAI} = \mathrm{ESR}_{c1}/\mathrm{ESR}_{c2} = P\left(Z \geq z^*_{c1}\right)/P\left(Z \geq z^*_{c2}\right). \tag{12}$$

In Equations 8 through 11, x^* is the value for X at y^* from the common regression line as depicted in Figures 3 and 4. The slope of the common regression line is given by

$$\beta = \rho\frac{\sigma_Y}{\sigma_X} \tag{13}$$

(e.g., Maxwell & Arvey, 1993, p. 434), and the y-intercept of the common regression line is

$$\alpha = \mu_Y - \beta\mu_X, \tag{14}$$

where $\mu_X, \sigma_X, \mu_Y, \sigma_Y$, and ρ are the population test score mean and standard deviation, performance mean and standard deviation, and test validity. In the special case of an unbiased test, $\beta = \beta_1 = \beta_2$ and $\alpha = \alpha_1 = \alpha_2$.

Alternatively, suppose that the group regression lines are not identical. To determine expected false positive and negatives, we need two additional Z-scores:

$$z^*_{g1} = \frac{x^*_1 - \mu_{X_1}}{\sigma_{X_1}} \tag{15}$$

$$z^*_{g2} = \frac{x^*_2 - \mu_{X_2}}{\sigma_{X_2}}, \tag{16}$$

where x_1^* and x_2^* are the expected selection cutoffs from the group-based lines at y^* (see Figure 4 and Equations B7 and B8). Hence, the "g" subscripts reference values computed at the group-based regression lines. If the test is biased, an expected (bias-based) false negative for Group 1 will occur when x^* is used to determine the expected selection cutoff and $z_{g1}{}^*$ $< z_{c1}{}^*$ (see Appendix A for details). Therefore, the probability of expected false negatives for Group 1 for the normal model will be

$$P\left(z_{g1}^* < Z < z_{c1}^*\right) \qquad \text{if} \quad z_{g1}^* < z_{c1}^*. \tag{17}$$

The probability of bias-based expected false positives for Group 1 is

$$P\left(z_{c1}^* < Z < z_{g1}^*\right) \quad \text{if} \quad z_{c1}^* < z_{g1}^*. \tag{18}$$

For Group 2, the probabilities of expected false negatives and positives are, respectively,

$$P\left(z_{g2}^* < Z < z_{c2}^*\right) \quad \text{if} \quad z_{g2}^* < z_{c2}^*, \quad \text{and} \tag{19}$$

$$P\left(z_{c2}^* < Z < z_{g2}^*\right) \quad \text{if} \quad z_{c2}^* < z_{g2}^*. \tag{20}$$

One important advantage of our framework is its ability to calculate numerical values for key concepts using straightforward mathematical relationships when normality is presumed. In particular, regression lines, expected performance levels, expected selection cutoffs, expected adverse impact, and probabilities of bias-based expected false positives and negatives are easily calculated using well-known regression relationships and with reference to a standard normal table of probabilities. However, we emphasize that our general framework is applicable to any stochastic specification and is not limited to those situations in which normality is present (see Appendix A). Furthermore, as described in Appendix A, our framework is readily generalizable to selection situations involving more than two groups as well as more than one predictor.

In the next two sections of the article, we illustrate the applicability and usefulness of our framework using three human resource selection scenarios. Scenario A, described in the section titled "Application I," refers to a situation in which the lines are identical across the two groups (i.e., the test is truly unbiased). Scenarios B and C, described in the section titled "Application II," refer to situations in which the lines are not identical across the two groups: In Scenario B, differences are based on intercepts, and in Scenario C differences are based on both intercepts and slopes. We chose Scenarios A and B for their likeness to actual selection situations as reported in the literature, thus providing a meaningful context for our work. Scenario C (i.e., differences in both intercepts and slopes) is not typically reported in the literature. However, we included this situation to illustrate the generalizability of our framework. To make our examples simple yet realistic, our three scenarios presume the use of a general mental abilities test (X) to predict performance (Y) as measured on a 5-point scale of supervisory ratings. We also assume that both groups' test scores and supervisory ratings follow a joint bivariate normal distribution (see Appendix B). Note that although many equations are involved in obtaining all the numerical values, the computations described in the next two sections are performed easily by using the online calculator that we describe in the Discussion section of this article.

Application I: Linking Desired Performance With Expected Adverse Impact (Unbiased Test)

In this application, we consider the desired performance–adverse impact tradeoff in relation to the 80% adverse impact benchmark, which has been institutionalized as a desirable target since the publication of the Uniform Guidelines on Employee Selection Procedures in 1978. Diversity is a goal of many employers that can be achieved best by selection procedures that produce similar proportions of qualified applicants. So, although our illustrations use the 80% rule of thumb as a desirable minimum target, our framework and online calculator allow for an examination of the consequences of using a particular test in relation to any adverse impact proportion.

Scenario A: An Unbiased Test

In Scenario A, we set the mean test score for Group 2 (i.e., majority group) at 100 ($\mu_{X_2} = 100$) and at 92.8 for Group 1 ($\mu_{X_1} = 92.8$). This is consistent with differences between mean scores for African Americans and Whites reported in the literature (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Because the difference in general mental ability mean scores between groups varies based on setting, sample, and type of construct assessed (e.g., fluid vs. crystallized intelligence; Hough et al., 2001), we are using these specific values as mere illustrations. We set the standard deviations equal for both groups (i.e., $\sigma_{X_1} = \sigma_{X_2} = 10$ and $\sigma_{Y_1} = \sigma_{Y_2} = 1$) and, consistent with previous findings, presume that the test is equally valid for both groups ($\rho_1 = \rho_2 = .5$; cf. Hunter, Schmidt, & Hunter, 1979). The mean supervisory rating is set at 3.11 for Group 2 ($\mu_{Y_2} = 3.11$) and 2.75 for Group 1 ($\mu_{Y_1} = 2.75$), which are consistent with results published recently regarding mean standardized differences in job performance between African Americans and Whites (Roth, Huffcutt, & Bobko, 2003). We also presume that $\mu_X = 98.56, \sigma_X = 10.406, \mu_Y = 3.038, \sigma_Y = 1.0103$, and $\rho = .515$. Although these parameter values are derived from the group-specific parameters using the assumptions that there are only two groups and that Group 1 candidates comprise 20% of the total population, our general formulation is not restricted to these two conditions. Collectively, these parameters coincide with an unbiased test because, from Equations 6 and 7 (and their analogues for Group 2), and Equations 13 and 14, $\beta_1 = \beta_2 = \beta = .05$ and $\alpha_1 = \alpha_2 = \alpha = -1.89$.

Suppose that the desired performance level is 3.25 on the 5-point scale. At $y^* = 3.25$, the expected selection cutoff is $x^* = 102.8$ (Equation B6). The expected selection ratio for Group 1 applicants is 15.87% (Equations 8 and 9); the expected selection ratio for Group 2 applicants is 38.97% (Equations 10 and 11). Finally, expected adverse impact is 40.7% (Equation 12), well below the 80% benchmark considered satisfactory by the Uniform Guidelines.

Figure 5 shows the relationship between desired performance levels and expected adverse impact for Scenario A. To derive the values plotted in Figure 5, we varied y^* from zero to five and used Equation 12 to compute the corresponding expected adverse impact for each value of y^*. Superimposed on this graph is the 80% adverse impact benchmark. To just reach the 80% benchmark using this particular test, Figure 5 shows that the organization must lower the desired performance level from 3.25 to 2.45. At $y^* = 2.45$, the expected selection cutoff is $x^* = 86.8$, which produces expected selection ratios of 72.6% for Group 1 and 90.7% for Group 2. Thus, in this particular scenario, to reach the 80% benchmark, an organization would expect to select large percentages from both populations.

In short, our integrative framework provides a method for directly linking desired performance levels with expected selection ratios. Knowing the expected selection ratios allows for the computation of expected adverse impact. In this particular scenario, we conclude that although this is an unbiased test and there is validity evidence, an organization may choose not to use this particular test because in order to avoid substantial expected adverse impact it would have to set predicted performance levels much lower than desired and, thus, expect to hire a very large proportion of applicants from both groups.

Application II: Expected Selection Errors That Arise From Incorrectly Believing That a Test Is Unbiased

Since the passage of the Civil Rights Act of 1991, the use of differential selection cutoff scores and group-based regression lines is unlawful. In other words, organizations must use the same regression equation and selection cutoffs with all applicants regardless of group membership.

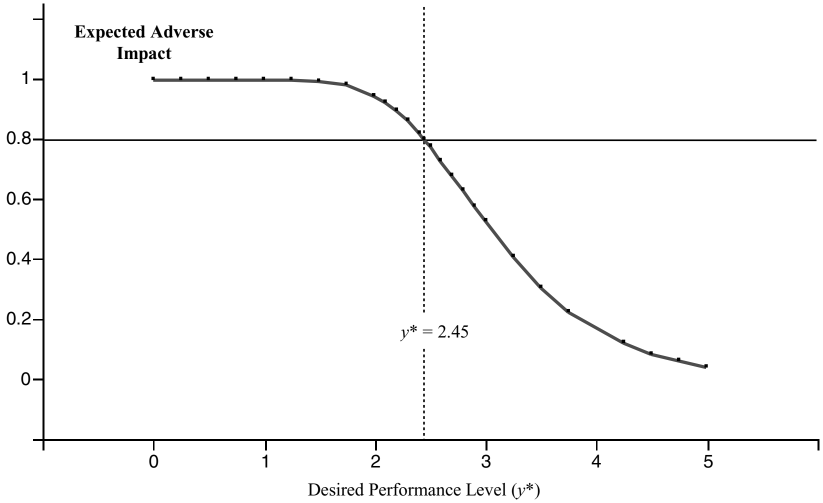


Figure 5: **Relationship between Desired Performance Levels (y^*) and Expected Adverse Impact for an Unbiased Test (Scenario A).**

As noted in the Introduction section, to determine whether a test is unbiased, researchers typically use a multiple regression framework in which race, sex, and other categorical variables related to protected class status are entered as moderators (AERA, APA, & NCME, 1999, Standard 7.6; Campbell, 1996; Cleary, 1968; Hough et al., 2001). Unfortunately, several Monte Carlo simulations (e.g., Aguinis, Boik, & Pierce, 2001; Aguinis & Stone-Romero, 1997) demonstrated that the moderator test has very low statistical power. One conclusion from this body of research is that very large samples are needed to detect differences in slopes across groups even when large differences exist in the populations. Indeed, Aguinis and Stone-Romero (1997) issued the warning that due to the low power of the test bias assessment procedure "practitioners may inappropriately use personnel selection tests that predict performance differentially for various subgroups" (p. 203). More recently, Aguinis et al. (2005) suggested that "past null findings be closely scrutinized to assess whether they may have been due to the impact of artifacts as opposed to the absence of a moderating effect in the population" (p. 101). Put another way, in many situations, organizations believing that they are in compliance with the Civil Rights Act of 1991 might unknowingly be using a biased test as if it were unbiased. In these situations, using our framework reveals that organizations will face unanticipated bias-based expected false positives and false negatives, as well as unanticipated performance levels from both groups, as demonstrated by Scenario B.

Scenario B: A Biased Test Believed to be Unbiased (Intercept Differences)

Scenario B uses the same group-specific parameters as in Scenario A except we set $\mu_{Y_2} = 3.5$. These parameters coincide with a biased test characterized by different yet parallel regression lines:

$$\text{Group 1: } E(Y_1 \mid X_1) = -1.89 + 0.05X_1$$

$$\text{Group 2: } E(Y_2 \mid X_2) = -1.5 + 0.05X_2.$$

In other words, differences in regression equations between groups are due to differences in intercepts, which is a common finding (e.g., Hunter & Schmidt, 1976; Reilly, 1973; Rotundo & Sackett, 1999). For Scenario B, an individual from Group 1 is expected to perform .39 points lower on average on the 5-point performance scale than an individual from Group 2 with the same test score. Referring back to Figure 4, the vertical distance between the group-based regression lines is .39 points. Alternatively, at any given performance level, Group 1's expected selection cutoff will be 7.8 points higher than that for Group 2 because the distance between x_1^* and x_2^* at any chosen y^* is 7.8. We also presume in Scenario B that $\mu_X = 98.56$, $\sigma_X = 10.41$, $\mu_Y = 3.35$, $\sigma_Y = 1.04$, and $\rho = .54$. Using Equations 13 and 14, the common regression line for Scenario B is as follows:

$$\text{Common Regression Line: } E(Y \mid X) = -1.967 + 0.053948X$$

Suppose that a particular organization wishes to hire individuals who are able to perform at a minimum level of three points on the 5-point supervisory rating scale ($y^* = 3$). Let's say that decision makers conduct the usual moderator test or related analyses (i.e., Lautenschlager & Mendoza, 1986) and conclude that the test is unbiased and, consequently, use the common regression line to choose an expected selection cutoff of x^*. The common regression line predicts an expected selection cutoff score on the general mental abilities test of 92.07 for both groups (Equation B6) and, therefore, at $y^* = 3$, expected adverse impact is 67.3% (Equation 12). Let's further assume that, contrary to the null statistical significance result regarding test bias, the test is actually biased (i.e., there are intercept-based differences between the regression lines). Under this scenario, the probability of expected false negatives is 5.5% of Group 2 applicants (Equation 19) and the probability of expected false positives is 22% of applicants from Group 1 (Equation 18). Put another way, as shown in Figure 6 at $y^* = 3$, making decisions based on the incorrect presumption of lack of test bias will result in failing to hire 5.5% of qualified Group 2 applicants and in hiring 22% of Group 1 applicants who are not qualified.

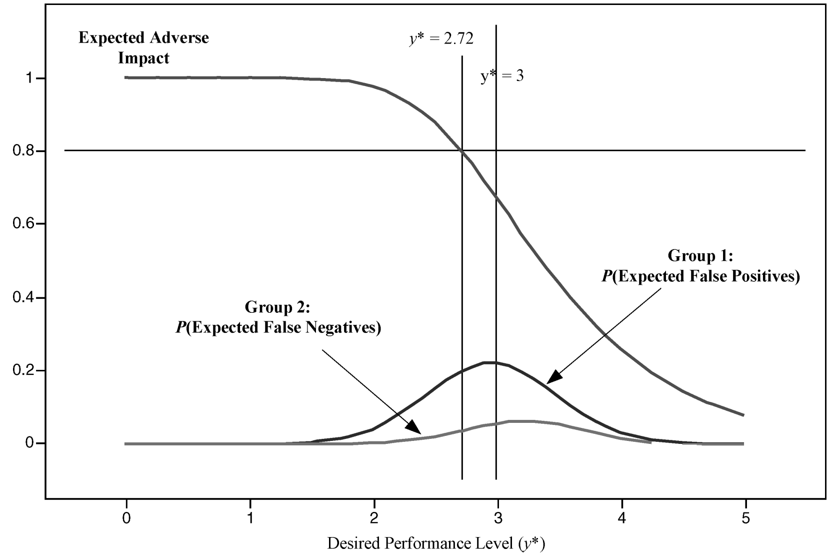


Figure 6: **Relationships Between Desired Performance Levels (y^*), Expected Adverse Impact, and Probabilities of Expected False Negatives and Expected False Positives for a Biased Test (Based on Intercept Differences) Believed to be Unbiased (Scenario B).** Group 1 has no expected false negatives over the range displayed in this graph. Group 2 will have expected false positives once y^* exceeds 4.42, but these values are miniscule (<.001).

As in Scenario A, we can obtain the precise value for y^* that just meets the 80% adverse impact benchmark by varying y^* and calculating expected adverse impact from Equation 12. Figure 6 illustrates the result of this analysis. To attain the 80% benchmark, the organization must lower its desired performance level from 3.00 to 2.72, where the expected selection cutoff from the common regression line is $x^* = 86.88$ and expected adverse impact is 79.9%. At $y^* = 2.72$ and $x^* = 86.88$, the organization expects to select 72.3% of the Group 1 applicants and 90.5% of those from Group 2. Furthermore, our analysis indicates that at $x^* = 86.88$ (the value associated with the 80% benchmark and the common regression line), 19.9% of individuals from Group 1 (the minority group) will not meet the expected performance standard and 3.5% of qualified individuals from Group 2 (the majority group) could meet the $y^* = 2.72$ performance level but are not under consideration for employment (see Figure 6 at $y^* = 2.72$). Similar to Scenario A, to meet the 80% adverse impact benchmark, the organization must lower its desired performance level to the point where large percentages of both populations are under consideration for employment.

Furthermore, at $x^* = 86.88$, average performance of both groups will deviate from the desired performance level predicted by the common regression line (cf. Figure 1). Although the expected performance for both groups is 2.72 via the common regression line, in this case in which the test is actually biased, the true average performance levels will be 2.84 for Group 2 and 2.45 for Group 1 (Equations B10 and B11). Group 2 will perform .12 points better than expected on average, but Group 1 will perform .27 points worse than expected on average.

Scenario C: A Biased Test Believed to be Unbiased (Intercept and Slope Differences)

Scenario C's group parameters are identical to those in Scenario B except we increase the standard deviation of test scores for Group 1 to $\sigma_{X_1} = 15$. Consequently, the regression line for Group 2 is steeper than that for Group 1; that is, in Scenario C, the group-based regression lines differ regarding both intercepts and slopes. The regression line for Group 2 is the same as in Scenario B. Group 1's regression line for Scenario C is:

$$\text{Group 1 Regression Line}: E(Y_1 \mid X_1) = -0.3433 + 0.033X_1.$$

For Scenario C, we set $\mu_X = 98.56$, $\sigma_X = 11.5$, $\mu_Y = 3.35$, $\sigma_Y = 1.04$, and $\rho = .53$.

Figure 7 includes a plot of expected adverse impact against the desired performance level for Scenario C. The relationship between expected adverse impact and y^* is nonmonotonic. For small values of y^*, virtually everyone is under consideration for selection from both groups, so expected adverse impact is close to a highly desirable value of 1.0. As y^* increases, the expected pool of eligible Group 1 applicants declines at a faster rate than that of Group 2 and expected adverse impact reaches undesirable levels. It eventually increases as the expected hiring pool for Group 2 declines faster than the expected hiring pool of Group 1. Expected adverse impact can exceed 1.0 because $\sigma_{X_1} > \sigma_{X_2}$, which means that the tail area of $f(X_1)$ will eventually exceed that of $f(X_2)$ for large values of y^*.

Interestingly, Figure 7 shows that there are two values for y^* that will just meet the 80% benchmark for Scenario C using the common regression line: at $y^* = 2.53$ and at $y^* = 3.91$. If the organization's desired performance level falls within the range of 2.53 to 3.91, expected adverse impact will be less than the 80% benchmark. Only those desired performance levels in the high and low ranges coincide with meeting or exceeding the 80% benchmark.

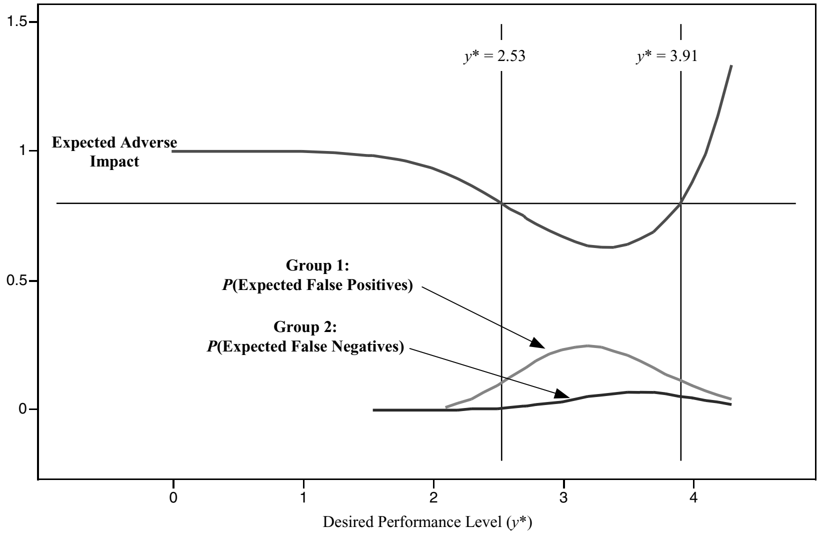


Figure 7: **Relationships Between Desired Performance Levels (y^*), Expected Adverse Impact, and Probabilities of Expected False Negatives and Expected False Positives for a Biased Test (Based on Intercept and Slope Differences) Believed to be Unbiased (Scenario C).** Group 1 expected false negatives and Group 2 expected false positives are so small that they are undetectable in this graph.

At the desired performance level of $y^* = 3.91$, the relatively high expected selection cutoff ($x^* = 110.24$) calls for an expected selection ratio of 12.2% for Group 1. However, because in this scenario the test is actually biased, the vast majority of those people under consideration for selection from Group 1 will fail to meet expectations of $y^* = 3.91$ because the probability of expected false positives is 11.2% for Group 1 (Figure 7). For Group 2, 20.6% are expected to be able to perform at or above $y^* = 3.91$, but only 15.3% are under consideration for employment (i.e., the probability of a false negative for Group 2 is 5.3%). Furthermore, at $x^* = 110.24$ and $y^* = 3.91$, the average performance for Group 1 will be less than, and that for Group 2 greater than, the desired level because $y^* = 3.91 > y_1^* = 3.33$ and $y^* = 3.91 < y_2^* = 4.01$.

At the other end of the spectrum, the expected selection cutoff at $y^* = 2.53$ is $x^* = 81.45$. The expected selection ratios are 77.5% for Group 1 and 96.8% for Group 2. Group 1 will perform .16 points lower than expected on average ($y_1^* = 2.37$) and have a 10.5% expected false positive rate. Group 2 will perform very close to expectations ($y_2^* = 2.57$) and will have a negligible (.6%) rate of expected false negatives. Accordingly, just meeting the 80% expected adverse impact target can be achieved with little compromise in expected performance for Group 2 and virtually no bias-based selection error for Group 2. Unfortunately, virtually all of Group 2 would be in expected hiring pool.

Discussion

This article addresses a void in the literature regarding relationships among the key and interrelated concepts of (a) test validity, (b) test bias, (c) selection errors, and (d) adverse impact. We proposed a framework based on statistical principles that integrates these four concepts into one comprehensive planning tool and allows researchers and practitioners to assess numerically how the four concepts interact with each other. To make our approach more user friendly and accessible, we have designed a computer program written in Java that is available for free and can be executed online at http://www.cudenver.edu/∼haguinis (click on "Selection Program" on the left). This computer program assumes bivariate normality for both groups and performs all needed computations, including the creation of graphs similar to Figures 1 and 4 and an output table providing precise numerical values based on user-supplied input. The resulting graphs and tables can be used as an aid in decision making and analysis by researchers, test developers, employers, and policy makers.

Implications for Theory and Future Research

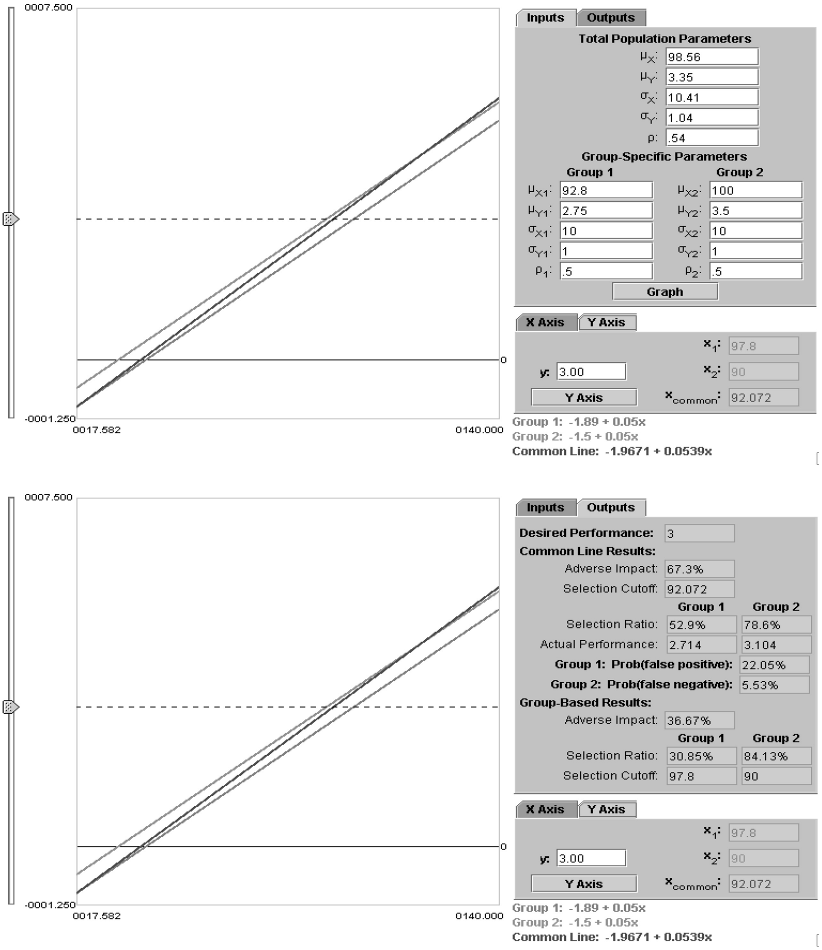
From a theory point of view, our framework provides a more complete picture of the selection process by integrating four key concepts that have not been examined simultaneously thus far. Because of this integration, our framework provides answers to several why-type questions such as the following: why various characteristics of the testing situation (e.g., test score means across groups, expected selection ratios) lead to expected adverse impact, why desired performance levels may need to be lowered in order to mitigate expected adverse impact, why there is a tradeoff between expected false positives and negatives across groups in some cases, why using a common regression line generates bias-based expected selection errors and unanticipated performance should a test prove to be biased, and so forth. We hope this integration will allow for fruitful areas of research in the future such as the development of selection tools that maximize validity, minimize bias and expected selection errors, and mitigate expected adverse impact.

Implications for Practice

Our framework provides test developers, employers, and the legal system with a broader perspective regarding practical consequences associated with various selection systems that vary regarding their degree of validity and bias. Test developers and employers are mostly concerned about selection accuracy whereas policy makers are concerned about accuracy but are also concerned about broader societal issues (Cascio, Goldstein, Outtz, & Zedeck, 2004). Our framework is sufficiently broad to allow each of these stakeholders to answer key questions about human resource selection tests. For example, the implicit tradeoff between job performance and expected adverse impact and related workforce integration and diversity considerations can be considered explicitly. Decision makers can thus combine psychometric with other important value-based considerations before using selection tests.

To use our framework, we propose the following process that is easily implemented using the computer program mentioned above. First, input the mean test score, mean performance score, test score standard deviation, performance standard deviation, and validity coefficient for each group and for the population as a whole (i.e., all individuals combined regardless of group membership). The group with the lowest mean test score should always be labeled as Group 1. An illustrative input screen using the parameters described above for Scenario B is included in Figure 8 (top panel). As shown in Figure 8 (top panel), the program will graph the three regression lines: (a) common regression line, (b) regression line for Group 1, and (c) regression line for Group 2 (similar to Figure 1). The program will graph all three lines even if the usual statistical tools that are used to assess potential test bias (e.g., Aguinis, 2004; Lautenschlager & Mendoza, 1986) show no statistically significant differences in the group-based intercepts and/or slopes. The lines will be identical in the display only if the input is such that the intercepts and the slopes are exactly equal for each group and the population as a whole.

After supplying the parameters to the input screen, clicking on the "Outputs" tab produces an output screen such as the one shown in Figure 8 (bottom panel). The user can supply the program with the desired performance level (shown to be 3.0 in Figure 8's bottom panel). The output screen will then display the associated expected adverse impact (67.3% in Figure 8's bottom panel), expected selection cutoff (92.072), group-specific expected selection ratios and expected performance levels (e.g., 52.9% and 2.714 for Group 1), and probabilities of expected bias-based selection errors. Should any of these outcomes be undesirable to decision makers, sensitivity analysis can be performed by varying y^* until key outcomes such as expected adverse impact or desired performance levels



Total Population Parameters

μ_X: 98.56
μ_Y: 3.35
σ_X: 10.41
σ_Y: 1.04
ρ: .54

Group-Specific Parameters

Group 1		Group 2	
μ_{X1}:	92.8	μ_{X2}:	100
μ_{Y1}:	2.75	μ_{Y2}:	3.5
σ_{X1}:	10	σ_{X2}:	10
σ_{Y1}:	1	σ_{Y2}:	1
ρ_1:	.5	ρ_2:	.5

x_1: 97.8
y: 3.00
x_2: 90
Y Axis
x_{common}: 92.072

Group 1: -1.89 + 0.05x
Group 2: -1.5 + 0.05x
Common Line: -1.9671 + 0.0539x

Desired Performance: 3
Common Line Results:
Adverse Impact: 67.3%
Selection Cutoff: 92.072

	Group 1	Group 2
Selection Ratio:	52.9%	78.6%
Actual Performance:	2.714	3.104

Group 1: Prob(false positive): 22.05%
Group 2: Prob(false negative): 5.53%

Group-Based Results:
Adverse Impact: 36.67%

	Group 1	Group 2
Selection Ratio:	30.85%	84.13%
Selection Cutoff:	97.8	90

x_1: 97.8
y: 3.00
x_2: 90
Y Axis
x_{common}: 92.072

Group 1: -1.89 + 0.05x
Group 2: -1.5 + 0.05x
Common Line: -1.9671 + 0.0539x

Figure 8: **Input (Top Panel) and Output (Bottom Panel) Screens for Computer Program That Implements all Required Calculations.**

reach acceptable levels. This can be easily done by holding, clicking, and moving the slider that appears on the "Y axis" of the output screen. Examination of the numerical values on the output screen associated with different *y** values will determine what effect the differing desired performance levels will have on expected performance, expected selection ratios and adverse impact, and bias-based expected selection errors. The user is then in a position to determine whether these outcomes would be acceptable and, therefore, whether the test would be used.

The application of our integrative framework to actual tests in actual selection contexts allows test developers and employers to understand selection decision consequences before a test is put to use. Following the procedure outlined above allows for an estimation of practically meaningful consequences (e.g., expected selection errors and expected adverse impact) of using a particular test *regardless of the results of the test bias assessment*. Thus, our framework allows for an understanding of the practical significance of potential test bias regardless of the statistical significance results, which often lead to Type II errors (e.g., Aguinis, 1995, 2004; Aguinis et al., 2005; Aguinis & Stone-Romero, 1997). In other words, our framework does not rely on null hypothesis significance testing, which has been criticized heavily on numerous grounds (e.g., Cashen & Geiger, 2004; Cortina & Folger, 1998).

Finally, note that some users may utilize input values based on statistics derived from small samples. These statistics (e.g., means and validity coefficients for each group) are the best estimators of their respective parameters, but they are influenced by sampling error (Aguinis, 2001). Thus, when input values are based on small sample sizes, computations using the program can be made using ranges of values that fall within each statistic's confidence interval in addition to the point estimates.

Implications for Policy Making

Important new insights and public policy implications arise from the use of our integrative framework regarding the use of test scores as mandated by the Civil Rights Act of 1991. For example, return briefly to the group-based parameters from Scenario B, but set the common regression line to $E(Y \mid X) = -6.91667 + .10417X$. In this scenario, an examination of the values for y^* that coincide with expected adverse impact of at least 80% in Figure 9 indicate that, over this range and for any given y^* value, there is less expected adverse impact when group-based regression lines are used than when the common regression line is used (Appendix C details the calculations needed to produce the values plotted in Figure 9). Our framework allows for the conclusion that although the Civil Rights Act of 1991 prohibits differential selection cutoffs, such a prohibition means that in some situations expected adverse impact becomes more severe (as compared to using group-based lines). On the other hand, in other instances, using a common regression line and one selection cutoff for both groups (regardless of whether the test is actually biased or unbiased) can lead to less severe expected adverse impact. This phenomenon is shown in Figure 9 when $y^* \geq 2.5$. Note, however, that this range of values coincides with expected adverse impact values smaller than 80%.

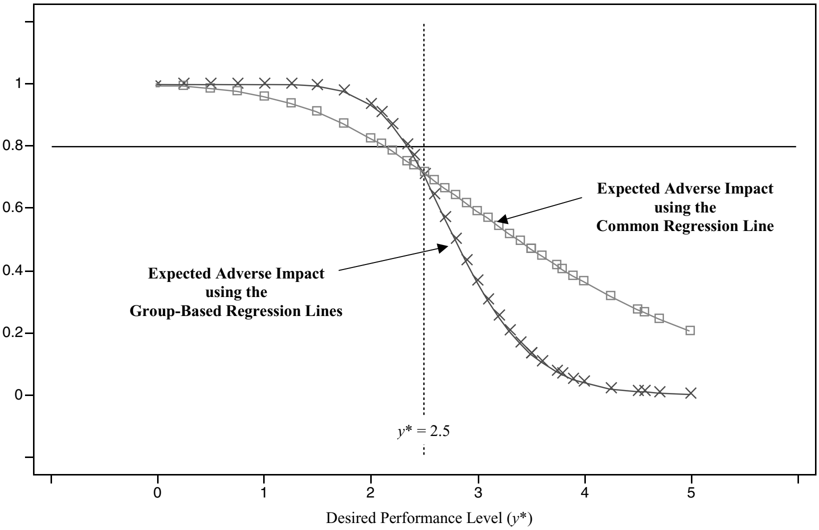


Figure 9: **Relationships Between Desired Performance Level (y^*), Expected Adverse Impact for a Test Believed to be Unbiased, and Expected Adverse Impact for a Biased Test.**

Given the Civil Rights Act of 1991, the use of group-based regression lines and cutoff scores is not legally permissible. However, if the intent of the Act is to not discriminate against members of protected classes and to mitigate adverse impact and its consequences, then our framework provides a powerful tool that could be used to explore situations in which the use of group-based lines and expected cutoff scores may be congruent with the Act's intent. We are not advocating the generally unlawful practice of using group-based regression lines and differential expected cutoff scores. As noted by an anonymous reviewer, given that many studies do not have enough power to detect test bias, it would be hard to justify establishing group-based cut scores particularly with a variable like race that is not sufficiently discrete (i.e., many people are mixtures of different races and could legitimately choose to belong to whichever group has the lower cut score). However, our framework can be used as a tool to help inform future policy making regarding situations in which public policy may lead to desirable, and undesirable, outcomes.

Underlying Assumptions and Potential Limitations

We note six underlying assumptions and potential limitations of our integrative framework. First, as is the case when a validity coefficient is used to make decisions regarding a test, our framework assumes that the criterion measure is not biased. In general, there is a consensus in the human resource selection literature that supervisory ratings are free from racial bias (e.g., Cascio & Aguinis, 2005a; Waldman & Avolio, 1991). However, a recent study by Stauffer and Buckley (2005), which reanalyzed data previously collected by Sackett and DuBois (1991), concluded that "if you are a White ratee, then it does not matter whether your supervisor is Black or White. If you are a Black ratee, then it is important whether your supervisor is Black or White" (p. 589). The preponderance of the evidence thus far is in favor of the no-bias conclusion. However, in light of Stauffer and Buckley's conclusions, we acknowledge the underlying assumption in our framework that performance data are unbiased.

Second, an underlying assumption in the use of individual-level criterion data in computing the validity coefficient is that the primary goal of the selection system is the maximization of individual performance. However, organizations may wish to maximize team performance or maximize organizational effectiveness, which may depend less on individual performance and more on unit- and team-level performance. In spite of this underlying assumption in using validity coefficients, our framework does allow for the maximization of other, and sometimes competing, goals. In fact, our framework allows for an explicit consideration of tradeoffs involved in using a particular test to maximize job performance measured at the individual level in relation to other goals at the unit or organizational level (i.e., mitigation of expected adverse impact). So, our framework allows for the consideration of both objective individual-level and higher-level concerns and as well as both psychometric and value-based factors in using tests and therefore may allow test developers and users to reach a "cultural optimum" (Darlington, 1971) in which both psychometric and other value-based principles are considered (Zedeck & Goldstein, 2000). Thus, we do not see the use of individual-level criterion data as a limitation of our framework.

Third, each of the three scenarios we presented to illustrate the applicability of our integrated framework presumes that meeting the 80% expected adverse impact benchmark is of primary concern to organizations. As noted earlier, however, our framework and online calculator allows for analyses based on any targeted expected adverse impact proportion or, for that matter, using any other criterion as the primary target of focus (e.g., minimizing expected bias-based selection errors).

Fourth, as has been common practice in the personnel selection literature for over 50 years, we make the assumption that "the applicant group and the present employee group are similarly constituted" (Taylor & Russell, 1939, p 567). Thus, our framework and calculations apply to the extent that important differences do not exist between the individuals used to obtain total population and group-based parameter estimates and future applicants.

Fifth, our illustrative Scenarios A through C and the online calculator presume bivariate normality. To the extent that a particular situation is known to deviate strongly from normality, the results from our online calculator should be considered only approximate. Note, however, that the normality assumption used in a portion of this article is not a limitation of our general framework. Indeed, the general framework proposed in Appendix A is applicable to any stochastic specification and does not presume any specific probability distribution. In short, nonnormality may affect the resulting numerical values (but not the conceptual framework). Future research could examine empirically the extent to which nonnormality may affect the resulting numerical values given various degrees of violation of the bivariate normality assumption.

Finally, our methodology and online calculator generate numerical estimates of bias-based selection errors (i.e., those caused by using a biased test as if it were unbiased) and not predictive selection errors (i.e., those caused by using less than perfect prediction systems). Readers interested in augmenting our framework to include predictive selection errors are encouraged to refer to Taylor and Russell (1939).

Concluding Remarks

Our integrative framework makes a contribution to theory and practice in that it allows for a better understanding of the relationship among four closely related issues in human resource selection: test validity, test bias, selection errors, and adverse impact. This integrated framework has the potential to lead to fruitful avenues of research regarding the intrinsic relationships among these key concepts. From a practical point of view, the proposed framework allows for a better assessment of selection outcomes before they actually take place and provides an informed evaluation of tradeoffs between expected performance, expected adverse impact, and expected selection errors regardless of whether moderated regression or other tools used to assess potential test bias indicate the test is biased. Finally, our framework can aid policy makers and the legal system because it allows for a better understanding of situations under which using differential selection rules across groups may be beneficial for, and harmful to, individuals, organizations, and society at large.

REFERENCES

Abramowitz M, Stegun IA. (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: Dover.

Aguinis H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management, 21*, 1141–1158.

Aguinis H. (2001). Estimation of sampling variance of correlations in meta-analysis. PERSONNEL PSYCHOLOGY, *54*, 569–590.

Aguinis H. (2004). *Regression analysis for categorical moderators*. New York: Guilford.

Aguinis H, Beaty JC, Boik RJ, Pierce CA. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*, 94–107.

Aguinis H, Boik RJ, Pierce CA. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods, 4*, 291–323.

Aguinis H, Stone-Romero EF. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192–206.

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

Biddle D. (2005). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Burlington, VT: Gower.

Bobko P, Roth PL. (2004). The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. In Martocchio J (Ed.), *Research in personnel and human resources management* (Vol. 19, pp. 177–197). New York: Elsevier.

Campbell JP. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behavior, 49*, 122–158.

Cascio WF, Aguinis H. (2005a). *Applied psychology in human resource management (6th edition)*. Upper Saddle River, NJ: Prentice Hall.

Cascio WF, Aguinis H. (2005b). Test development and use: New twists on old questions. *Human Resource Management, 44*, 219–235.

Cascio WF, Goldstein IL, Outtz J, Zedeck S. (2004). Social and technical issues in staffing decisions. In Aguinis H (Ed.), *Test-score banding in human resource selection: Legal, technical, and societal issues* (pp. 7–28). Westport, CT: Praeger.

Cashen LH, Geiger SW. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods, 7*, 151–167.

Civil Rights Act of 1991, 42 U.S.C. §§ 1981, 2000e et seq.

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

Cortina JM, Folger RG. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods, 1*, 334–350.

Curtis EW, Alf EF. (1969). Validity, predictive efficiency, and practical significance of selection tests. *Journal of Applied Psychology, 53*, 327–337.

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

Gatewood RD, Feild HS. (2001). *Human resource selection (5th edition)*. Stamford, CT: Harcourt.

Guion RM. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Hough LM, Oswald FL, Ployhart RE. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194.

Hunter JE, Schmidt FL. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin, 83*, 1053–1071.

Hunter JE, Schmidt FL, Hunter R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*, 721–735.

Hunter JE, Schmidt FL, Judiesch M. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*, 28–42.

Lautenschlager GJ, Mendoza JL. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10*, 133–139.

Lindgren BW. (1976). *Statistical theory (3rd edition)*. New York: MacMillan.

Martocchio JJ, Whitener EM. (1990). Fairness in personnel selection: A meta-analysis and policy implications. *Human Relations, 45*, 489–506.

Maxwell SE, Arvey RD. (1993). The search for predictors with high validity and low adverse impact: Compatible or incompatible goals? *Journal of Applied Psychology, 78*, 433–437.

Murphy KR, Shiarella AH. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. PERSONNEL PSYCHOLOGY*, 50*, 823–854.

Petersen NS, Novick MR. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3–29.

Ployhart RE, Schneider B, Schmitt N. (2006). *Staffing organizations: Contemporary practice and theory (3rd edition)*. Mahwah, NJ: Erlbaum.

Reilly RR. (1973). A note on minority group test bias studies. *Psychological Bulletin, 80*, 130–132.

Reilly RR, Chao GT. (1982). Validity and fairness of some alternative employee selection procedures. PERSONNEL PSYCHOLOGY*, 35*, 1–62.

Roth PL, Bevier CA, Bobko P, Switzer FS, Tyler P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. PERSONNEL PSYCHOLOGY*, 54*, 297–330.

Roth PL, Huffcutt AI, Bobko P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88*, 694–706.

Rotundo M, Sackett PR. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology, 84*, 815–822.

Sackett PR, DuBois CLZ. (1991). Rater–ratee race effects on performance evaluations: Challenging meta-analytic conclusions. *Journal of Applied Psychology, 76*, 873–877.

Schmidt FL, Hunter JE. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt FL, Pearlman K, Hunter JE. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. PERSONNEL PSYCHOLOGY*, 33*, 705–724.

Stauffer JM, Buckley MR. (2005). The existence and nature of racial bias in supervisory ratings. *Journal of Applied Psychology, 90*, 586–591.

Taylor HC, Russell JT. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565–578.

Thomas H. (1990). A likelihood-based model for validity generalization. *Journal of Applied Psychology, 75*, 13–20.

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

Uniform Guidelines on Employee Selection Procedures (1978). Federal Register, *43*, 38290–38315.

Waldman DA, Avolio BJ. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology, 76*, 897–901.

Zedeck S, Goldstein IL. (2000). The relationship between I/O psychology and public policy: A commentary. In Kehoe JF (Ed.) *Managing selection in changing organizations: Human resource strategies* (pp. 371–396). San Francisco: Jossey-Bass.

APPENDIX A
The General Formulation

We have a selection test (X) used to predict performance (Y). There are two groups for which we assume, without loss of generality (Roth, Bevier, Bobko, Switzer, & Tyler, 2001), that:

$$\mu_{X_1} \leq \mu_{X_2}, \quad \text{and} \quad \mu_{Y_1} \leq \mu_{Y_2}. \tag{A1}$$

Group 1 represents the minority group and Group 2 the majority group. Presume that the relationship between X and Y for each group can be represented by continuous bivariate distribution functions $f(X_1, Y_1)$ for Group 1 and $f(X_2, Y_2)$ for Group 2. A test is said to be unbiased if, for all $X = x$,

$$\mu_{Y_1 \mid X_1 = x} = \mu_{Y_2 \mid X_2 = x} \equiv h(x). \tag{A2}$$

That is, an unbiased test predicts the same mean performance for all individuals (regardless of group membership) who have the same test scores via the mean of the conditional distribution of Y given X. A biased test predicts different average performance for equivalent test scores.

In practice, the conditional mean function is also used to determine expected selection cutoffs. A desired performance level, y^*, is chosen. If the test is unbiased, then y^* is linked to test scores via Equation A2:

$$y^* = \mu_{Y_1 \mid X_1 = x^*} = \mu_{Y_2 \mid X_2 = x^*} = \mu_{Y \mid X = x^*} \equiv h(x^*), \tag{A3}$$

so that the *expected selection cutoff*, x^*, (again, if the test is unbiased) is given by:

$$x^* = h^{-1}(y^*), \tag{A4}$$

where x^* is the value for X predicted backwards through the conditional mean function at y^*. An individual is under consideration for selection when his or her score equals or exceeds x^*.

For the moment, consider Group 1. The expected selection ratio for Group 1 is given by:

$$P\left(X_1 \geq x^*\right) = \int_{x^*}^{\infty}\int_{-\infty}^{\infty} f(X_1, Y_1)\, dY_1 dX_1 = \int_{x^*}^{\infty}\int_{-\infty}^{\infty} f(X_1 \mid Y_1) f(Y_1)\, dY_1 dX_1$$

$$= \int_{x^*}^{\infty} f(X_1)\, dX_1 = 1 - F_{X_1}(x^*), \tag{A5}$$

where $f(X_1)$ is the marginal distribution function for X_1 and $F_{X_1}(.)$ is the cumulative distribution function of $f(X_1)$. Analogously, for Group 2, $P(\mathrm{X}_2 \geq x^*) = 1 - F_{\mathrm{X}_2}(x^*)$ is the expected selection ratio for Group 2 at (x^*, y^*). Therefore, expected adverse impact (EAI) for an unbiased test is:

$$\mathrm{EAI} = \frac{P(X_1 \geq x^*)}{P(X_2 \geq x^*)} = \frac{1 - F_{x_1}(x^*)}{1 - F_{x_2}(x^*)}. \tag{A6}$$

Our approach complements that of Maxwell and Arvey (1993), who used *d* as a measure of adverse impact. Our work extends theirs by noting that the expected selection ratio can be measured directly by referring to the marginal distribution function of the *X* variable.

When a test is biased, then

$$h_1(x^*) \equiv \mu_{Y_1 \mid X_1 = x^*} \neq \mu_{Y_2 \mid X_2 = x^*} \equiv h_2(x^*) \tag{A7}$$

for at least one x^*, so that expected selection cutoffs will differ by group:

$$x_1^* = h_1^{-1}(y^*) \tag{A8}$$

$$x_2^* = h_2^{-1}(y^*). \tag{A9}$$

If the test is biased, a (bias-based) expected false negative for Group 1 will occur when x^* from $h(x^*)$ is used to determine the expected selection cutoff and when $x_1^* < x^*$. The probability of expected false negatives for Group 1 is found by:

$$P\left(x_1^* \leq X_1 \leq x^*\right) = \int_{x_1^*}^{x^*}\int_{-\infty}^{\infty} f(X_1, Y_1)\, dY_1 dX_1 = F_{X_1}(x^*) - F_{X_1}(x_1^*) \tag{A10}$$

Analogously, a bias-based expected false positive for Group 1 occurs whenever $x^* < x_1^*$; its probability is $F_{X_1}(x_1^*) - F_{X_1}(x^*)$. For Group 2, probabilities of bias-based expected false negatives are $F_{X_2}(x^*) - F_{X_2}(x_2^*)$ when $x_2^* < x^*$ and expected false positives are $F_{\mathrm{X}_2}(x_2^*) - F_{\mathrm{X}_2}(x^*)$ when

$x^* < x_2^*$. With two groups, there are four possible combinations of bias-based expected false positives and negatives:

- Both groups will experience expected false negatives when $x_1^* < x^*$ and $x_2^* < x^*$ at a given y^*.
- Both groups will experience expected false positives when $x^* < x_1^*$ and $x^* < x_2^*$ at a given y^*.
- Group 1 will experience expected false positives and Group 2 expected false negatives when $x^* < x_1^*$ and $x_2^* < x^*$ at a given y^*.
- Group 1 will experience expected false negatives and Group 2 expected false positives when $x_1^* < x^*$ and $x^* < x_2^*$ at a given y^*.

Finally, at a given x^* value, y^* as shown in Figure 1 is derived using Equation A3. For group-based lines, $y_1^* = h_1(x^*)$ and $y_2^* = h_2(x^*)$ via Equation A7.

This general formulation makes no assumptions about the functional forms for $f(X_1, Y_1)$ or $f(X_2, Y_2)$, nor have we assumed that the conditional expectation of Y given X is linear or that the test is equally valid for both groups. It is generally applicable to any stochastic specification. In addition, our formulation readily accommodates more than two groups. Consider the situation in which there are k minority groups, say $1a$ through $1k$, with Group 2, once again, representing the majority group. In practice, expected adverse impact is calculated for the expressed purpose of comparing the minority (i.e., focal) to the majority (i.e., reference) group (Biddle, 2005). Therefore, expected selection cutoff, expected selection ratio, and expected adverse impact are calculated as above by replacing subscripts "1" with "1g" whenever the gth minority group is the focal group. For calculating probabilities of bias-based expected false positives and negatives in the presence of more than two groups, the common regression line shown in Figure 4 represents the regression line for all groups (i.e., $1a$ through $1k$ and 2) combined; in the notation above, the common regression line is simply $\mu_{Y|X=\mathrm{x}}$.

Finally, our formulation also applies to selection situations involving more than one assessment tool. Suppose, for example, that we are interested in calculating the expected selection ratio for Group 1 (or Group $1g$) using two tests, T_1 and T_2. The appropriate calculations for the expected selection ratio depend on how the organization chooses to use those two tests for selection purposes. We consider three possibilities.

1. The organization uses a linear combination of the two tests to form a composite test score, $\mathrm{T}_3 = \mathrm{a}_1\mathrm{T}_1 + \mathrm{a}_2\mathrm{T}_2$, with a_1 and a_2 being positive weights less than one and $\mathrm{a}_1 + \mathrm{a}_2 = 1$. In this case, in which a compensatory system is used, the mean, standard deviation, and correlation of the composite test can be derived using known formulas. Numerical

calculations for expected selection cutoffs, expected selection ratios, and so forth follow from above after replacing X with T_3.

2. The organization defines the expected hiring pool to be those individuals who score at least t_1^* on test T_1 *and* at least t_2^* on test T_2. In this case, the expected selection ratio is given by (suppressing group-based subscripts):

$$P\left(T_1 \geq t_1^* \text{ } and \text{ } T_2 \geq t_2^*\right) = \int_{t_1^*}^{\infty} \int_{t_2^*}^{\infty} f(T_1, T_2)\, dT_2 dT_1. \quad \text{(A11)}$$

3. The organization defines the expected hiring pool to include those individuals who score at least t_1^* on test T_1 *or* at least t_2^* on test T_2. Here, the expected selection ratio is

$$P\left(T_1 \geq t_1^*\right) + P\left(T_2 \geq t_2^*\right) - P\left(T_1 \geq t_1^* \text{ } and \text{ } T_2 \geq t_2^*\right). \quad \text{(A12)}$$

APPENDIX B
The Normal Model

In this appendix, we assume that the distribution functions $f(X_1, Y_1)$ and $f(X_2, Y_2)$ are bivariate normal with parameters $f(X_j, Y_j; \mu_{X_j}, \mu_{Y_j}, \sigma_{X_j}, \sigma_{Y_j}, \rho_j)$ for groups $j = 1, 2$.

Because $f(X_1, Y_1)$ is assumed bivariate normal, the conditional distribution of Y_1 given $X_1 = x$ is univariate normal with moments:

$$\mu_{Y_1 \mid X_1 = x} = \mu_{Y_1} + \rho_1 \frac{\sigma_{Y_1}}{\sigma_{X_1}}(x - \mu_{X_1}) \quad \text{(B1)}$$

$$\sigma^2_{Y_1 \mid X_1 = x} = \sigma^2_{Y_1}\left(1 - \rho_1^2\right) \quad \text{(B2)}$$

(e.g., Lindgren, 1976, p. 470). The assumption of bivariate normality implies that the conditional expectation (regression function) is linear in X, as shown by Equation B1.

If the test is truly unbiased, then:

$$\mu_{Y_1} + \rho_1 \frac{\sigma_{Y_1}}{\sigma_{X_1}}(x - \mu_{X_1}) = \mu_{Y_2} + \rho_2 \frac{\sigma_{Y_2}}{\sigma_{X_2}}(x - \mu_{X_2}). \quad \text{(B3)}$$

Equation B3 holds if and only if both groups have identical regression functions for all X; that is:

$$\rho_1 \frac{\sigma_{Y_1}}{\sigma_{X_1}} = \rho_2 \frac{\sigma_{Y_2}}{\sigma_{X_2}} = \beta \quad \text{(B4)}$$

$$\mu_{Y_1} - \beta\mu_{X_1} = \mu_{Y_2} - \beta\mu_{X_2} = \alpha. \quad \text{(B5)}$$

The expected selection cutoff for an unbiased test using the common regression line is calculated as:

$$x^* = (y^* - \alpha)/\beta. \tag{B6}$$

(Equation B6 equals Equation 1 when the test is unbiased because, for unbiased tests, $\alpha_1 = \alpha_2 = \alpha$ and $\beta_1 = \beta_2 = \beta$.)

If the test is biased, from Equation B1, $\beta_1 = \rho_1\ (\sigma_{Y1}/\sigma_{X1})$ and $\alpha_1 = \mu_{Y1} - \beta_1\ \mu_{X1}$ (and similarly for Group 2). Group-based expected selection cutoffs are a straightforward extension of Equation 1 using the group-based regression lines in Figure 4:

$$x_1^* = (y^* - \alpha_1)/\beta_1 \tag{B7}$$

$$x_2^* = (y^* - \alpha_2)/\beta_2. \tag{B8}$$

When *X* and *Y* are bivariate normal, the marginal distributions are univariate normal. Therefore expected selection ratios, expected adverse impact, and probabilities of bias-based expected false positives and expected false negatives involve standard normal probabilities as described in the body of the paper.

Finally, to compare differential performance scores at x^*, we use the following relationships (see Figure 1):

$$y^* = \alpha + \beta x^* \tag{B9}$$

$$y_1^* = \alpha_1 + \beta_1 x^* \tag{B10}$$

$$y_2^* = \alpha_2 + \beta_2 x^*. \tag{B11}$$

Our online calculator computes upper-tail probabilities of standard normal distributions using the 8-digit accuracy formula given by equation 26.2.16 in Abramowitz and Stegun (1965, p. 932).

APPENDIX C
Calculating Expected Adverse Impact for a Biased Test

We have previously introduced the concept of expected adverse impact for an unbiased test using the common regression line (see Figure 3 and Equation 12). Calculation of expected adverse impact when the regression lines differ across groups is a straightforward extension. In Figure 4, refer only to the group-based regression lines. If y^* is the desired performance level, a biased test produces expected selection cutoff x_1^* for Group 1 and x_2^* for Group 2. If an organization uses group-based selection cutoffs, applicants from Group 1 whose test scores exceed x_1^* comprise the Group 1 expected hiring pool and similarly for Group 2. Expected adverse impact for a biased test (EAI$_\text{B}$) is once again the ratio of the upper-tail areas of the marginal distributions of test scores, here at the group-based expected selection cutoffs, x_1^* and x_2^*:

$$\text{EAI}_\text{B} = P\left(X_1 \geq x_1^*\right) / P\left(X_2 \geq x_2^*\right) \tag{C1}$$

or assuming normality and via Equations 15 and 16,

$$\text{EAI}_\text{B} = P\left(Z > z_{g1}^*\right) / P\left(Z > z_{g2}^*\right). \tag{C2}$$

To find numerical values for the quantities outlined in this appendix, refer to the section of our online calculator output screen labeled, "Group-Based Results" (see Figure 8's bottom panel). When a biased test is used to find group-based expected selection cutoffs (and if the test is truly biased), there will be no expected biased-based errors; furthermore, expected performance is accurately predicted for both groups—at y^* (see Figure 4). Therefore, the "Group-Based Results" section of the output screen displays no probabilities for bias-based expected false positives and/or negatives. Those values (and differential performance predictions) only arise in our framework when a biased test is used as if it were unbiased.