Adverse Impact

Implications for Organizational Staffing and High Stakes Selection

Edited by

James L. Outtz

Routledge
Taylor & Francis Group
New York London

Routledge
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Routledge
Taylor & Francis Group
27 Church Road
Hove, East Sussex BN3 2FA

Routledge is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-0-8058-6374-1 (Hardback)

Library of Congress Cataloging-in-Publication Data

Adverse impact : implications for organizational staffing and high stakes selection / edited by James Outtz.
p. cm. -- (SIOP organizational frontiers series)
Includes bibliographical references and index.
ISBN 978-0-8058-6374-1 (alk. paper)
1. Employee selection. 2. Employment tests. 3. Personnel management. 4. Psychology, Industrial. I. Outtz, James. II. Society for Industrial and Organizational Psychology (U.S.).

HF5549.5.S38A385 2010
658.3'112--dc22
2009003530

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the Psychology Press Web site at
http://www.psypress.com

15

Balancing Adverse Impact, Selection Errors, and Employee Performance in the Presence of Test Bias

Herman Aguinis and Marlene A. Smith

Introduction

Adverse impact (AI) is a central issue in organizational staffing and high-stakes selection. Although this concept has a long history (Zedeck, 2009), it is usually operationalized as a ratio of two selection ratios (SRs) (Biddle, 2005; Bobko & Roth, 2004). AI = 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 for AI to be as close to 1.0 as possible because AI = 1.0 means that the selection ratios are identical across groups (e.g., ethnic majority and ethnic minority groups). However, the 80% AI benchmark (i.e., AI = 0.80) has been institutionalized as a desirable target since the publication of the *Uniform Guidelines on Employee Selection Procedures* in 1978. Specifically, Section A notes that "a selection rate for any race, sex, or ethnic group which is less than 4/5ths (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact" (p. 38297). Federal agencies use the 80% benchmark when judging compliance with federal guidelines. For example, Roth, Bobko, and Switzer (2006) noted that the typical first step in compliance proceedings includes checking the 80% benchmark and continuing with the process only if this benchmark is not met. Violating the 80% benchmark has important and often very costly implications for organizations, and in most situations, organizations will be better off avoiding

or at least mitigating AI. In practice, this means that personnel selection decision makers try to achieve an AI ratio of at least 0.80.

Achieving an acceptable AI ratio (i.e., AI ≥ 0.80) is often difficult when measuring constructs such as general mental abilities (GMAs), which are known to result in mean score differences across ethnicity-based groups (Aguinis, 2004b). Accordingly, personnel selection decision makers are often faced with a paradoxical situation: Using GMA and other predictors that maximize individual performance and resulting economic utility, as is typically conceptualized in human resources management and industrial and organizational psychology (Cascio & Aguinis, 2005, Chapter 3), often leads to the exclusion of members of ethnic minorities (Aguinis, Cortina, & Goldberg, 1998; Murphy, 2004).

Test bias exists when the same test score leads to different predicted performance scores for members of groups based on protected class status (e.g., race, sex). The presence of 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 (Aguinis, 2004a; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 1999, Standard 7.6; Cleary, 1968; Hough, Oswald, & Ployhart, 2001). Assessing test bias often leads to the incorrect conclusion that there is no bias because of low statistical power (Aguinis, Beaty, Boik, & Pierce, 2005; Aguinis, Boik, & Pierce, 2001; Aguinis & Stone-Romero, 1997). In other words, in many situations in which there is test bias, the test bias assessment procedures lead to the incorrect conclusion that bias is not present.

In this chapter, we offer an expanded way of thinking about AI in organizational staffing and high-stakes selection. As we noted, extensive simulation studies have demonstrated that test bias often exists in spite of results that moderating effects by group are statistically nonsignificant (Aguinis et al., 2001, 2005; Aguinis & Stone-Romero, 1997; see Aguinis, 1995, 2004a, for reviews). As a result, the decision to mitigate AI by lowering selection cut scores leaves an important issue out of the picture (see Kehoe, 2009, for a detailed treatment of the relationship between cut scores and AI). What has been left out in previous treatments of the cut score-AI relationship is that lowering cut scores to reach more acceptable levels of AI must be weighed against the collateral damage due to test bias that often exists unbeknown to test developers and users: unexpected performance levels of individuals selected and unexpected bias-based selection errors (both false positives and false negatives). In this chapter, we use the Aguinis and Smith (2007) decision-making model and Web-based calculator to demonstrate why information about possible test bias should be brought *explicitly* into the decision-making process. By doing so, selection decision makers will have a more comprehensive picture of how changing cut

scores to mitigate AI can also influence the organization regarding other important outcomes: the performance of those individuals hired and bias-based false positive and false negative errors.

Basics Concepts and Terminology

In selection decision making, a test score random variable X and a job performance random variable Y are presumed to follow a joint probability distribution. Y is related to X via a regression line as shown in Figure 15.1. For simplicity, this figure includes two groups only; Group 1 represents the minority group (e.g., ethnic minority) and Group 2 the majority group (e.g., ethnic majority), but the model can be extended to multiple groups. Group 1 and Group 2 may follow a *common regression line* $E(Y|X) = \alpha + \beta X$. This common regression line represents an unbiased test because, at any given test score (x^* in Figure 15.1), it predicts identical performance levels y^* for both groups (AERA, APA, & NCME, 1999). This, of course, would be the ideal situation because no test bias exists. In other situations, however, and unbeknown to selection decision makers due to the low statistical power of the bias assessment procedures, each group may follow its unique *group-specific regression line*, which are also shown in Figure 15.1. If a test is biased, it will predict average performance $y_1^* = E(Y_1|x^*)$ for Group 1 and $y_2^* =$

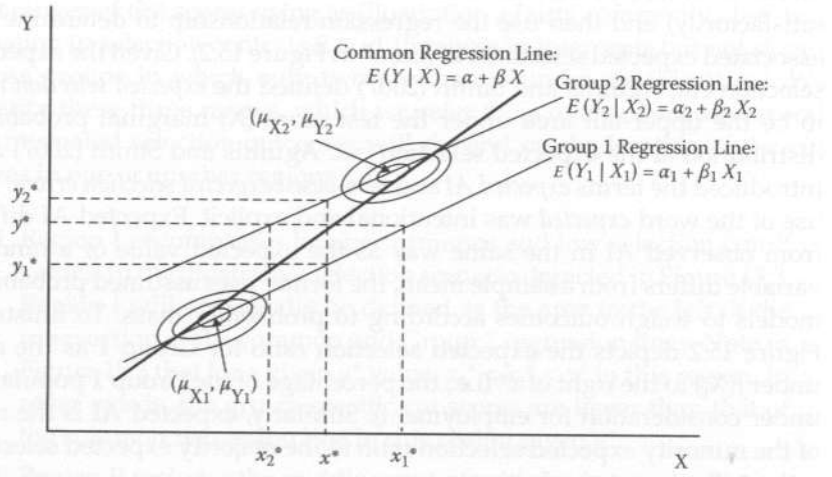

FIGURE 15.1
Common and group-specific regression lines and cut scores (Group 1 is the ethnic minority group).

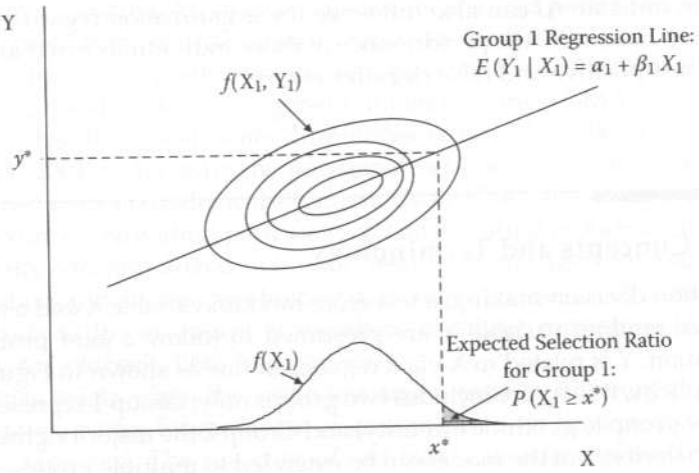

FIGURE 15.2
Expected selection ratio for Group 1 (i.e., ethnic minority group).

$E(Y_2|x^*)$ for Group 2 at test score x^*. The group-specific regression lines in Figure 15.1 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).

In many selection situations, decision makers stipulate a desired performance level (i.e., y^*, the minimum value for Y needed to perform the job satisfactorily) and then use the regression relationship to determine the associated expected selection cut (i.e., x^* in Figure 15.2). Given the expected selection cut, Aguinis and Smith (2007) defined the *expected selection ratio* to be the upper-tail area under the test score (X) marginal probability distribution at the expected selection cut. Aguinis and Smith (2007) also introduced the terms *expected AI* and *bias-based expected selection errors*. The use of the word *expected* was intentional and explicit. Expected AI differs from observed AI in the same way as the expected value of a random variable differs from a sample mean; the former uses assumed probability models to weigh outcomes according to probability mass. To illustrate, Figure 15.2 depicts the expected selection ratio for Group 1 as the area under $f(X_1)$ to the right of x^* (i.e., the percentage of the Group 1 population under consideration for employment). Similarly, expected AI is the ratio of the minority expected selection ratio to the majority expected selection ratio at the expected selection cut.

Although it may be small in magnitude, test bias exists every time that the group-specific lines do not overlap perfectly. When test bias exists, there are three possible cut scores associated with performance level y^*

(see Figure 15.1): (a) one to be used for both groups based on the common regression line (i.e., x^*), (b) one to be used for Group 1 based on its group-specific line (i.e., x_1^*), and (c) one to be used for Group 2 based on its group-specific line (i.e., x_2^*). Since the passing of the Civil Rights Act of 1991, the use of group-specific lines and cut scores in selection decision making is generally unlawful. So, either because bias is not detected due to low statistical power or because it is generally unlawful to use differential cut scores, the common regression line is often used for both groups even when bias exists. In such situations, selection errors (i.e., bias-based expected selection errors) are inevitably introduced because using group-specific lines and cut scores would maximize decision-making accuracy. Therefore, considering test bias provides a more comprehensive picture and increases the complexity of the cut score–AI relationship in that different forms and the degree of bias will lead to different types of bias-based selection errors. Next, we discuss three ranges of test scores and conditions under which selection decision makers are likely to be surprised (in some cases quite unpleasantly) in terms of selection outcomes other than the AI they are attempting to mitigate by lowering cut scores.

Three Relevant Regions of Test Scores

Figure 15.3 includes a graphic display of what we identify as three important ranges of test scores using as illustration a fairly commonly observed situation in selection contexts (i.e., differences in intercepts but not slopes across groups in which only two groups are under consideration). We identify these three ranges, which we refer to as *regions*, because several unanticipated selection outcomes will depend on the location of the cut scores in one or another region:

I. Region I encompasses low-performance and low selection cutoff values. In the illustrative selection scenario depicted in Figure 15.3, Region I will specifically be defined as the area to the left of the intersection of the common and Group 1 regression lines. Note in Figure 15.3 that for a given y^* value, $x_2^* < x_1^* < x^*$ in this region. In other words, the group-specific cut scores are lower than that of the common regression line in this region given y^*.

II. Region II includes the middle range of performance and selection cutoffs. For a situation such as the one in Figure 15.3, this region includes the area between the intersection of the common and Group 1 regression lines and the intersection of the common and

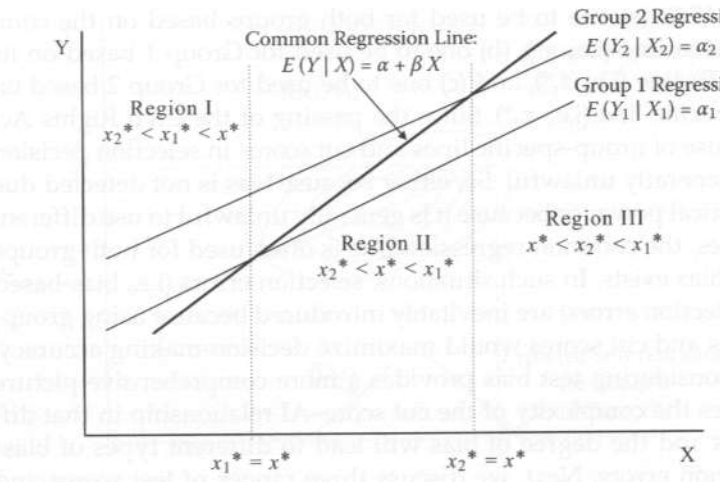

FIGURE 15.3
Three regions of test scores in the presence of intercept-based test bias.

Group 2 regression lines. In Figure 15.3, note that $x_2^* < x^* < x_1^*$ for any given performance level in this region; the common regression cut score lies between the group-specific cut scores.

III. Region III encompasses the high-performance, high selection cut score range. Referring again to Figure 15.3, this region is the area to the right of the intersection of the common and Group 2 regression lines so that $x^* < x_2^* < x_1^*$ for a given value of y^* (i.e., the common regression cut score is lower than that of the group-specific cut scores).

Understanding the Relationship Among Test Score Regions, Cut Scores, Expected Performance, Bias-Based Expected Selection Errors, and Expected Selection Ratios

In this section, we provide a discussion of what happens when test bias is present (albeit small in magnitude) and cut scores are lowered along the test score continuum to mitigate AI. We refer to the three regions identified and discuss implications in terms of (a) differentials between anticipated and actual performance of those individuals who are selected, (b) selectivity and utility of the selection system, and (c) bias-based selection errors (i.e., expected false positives and false negatives). To make our presentation more user friendly, we first keep our discussion general and use

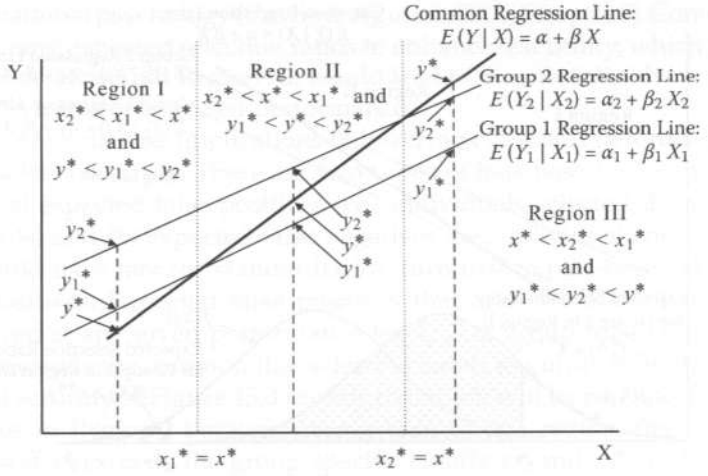

FIGURE 15.4
Performance differentials by test score region for intercept-based test bias.

graphs. We discuss two realistic numerical cases, including actual data, later in the chapter.

Consider again the illustrative and yet fairly typical situation in which there is bias based on intercept differences, but not slope differences, across two groups. Let us discuss first the issue of how those selected would perform relative to their anticipated performance level, as displayed in Figure 15.4. The severity and form of discrepancy between anticipated and actual performance depend on the region in which the cut scores are located. In Region I, selection decision makers would be pleasantly surprised because both groups would perform better than expected. That is because in Region I, decision makers, using the common regression line as mandated by law, expect performance level y^* for both groups. However, actual performance will be y_1^* for Group 1 and y_2^* for Group 2 because the test is biased and produces different performance levels for different groups. In Region II, results regarding performance are mixed. The majority group (Group 2) would perform better than expected on average, but the minority group would perform worse on average because for any given cut score in Region II, $y_1^* < y^* < y_2^*$. Finally, in Region III, unanticipated performance outcomes would be unpleasant all around: Both groups would perform worse than expected on average. Of course, results regarding each of the three regions would be accentuated to the extent that bias is more severe.

Consider now the implications of changing cut scores to mitigate AI in terms of the degree of selectivity of the system. As depicted in Figure 15.5, expected selection ratios in Region I will be larger than expected selection

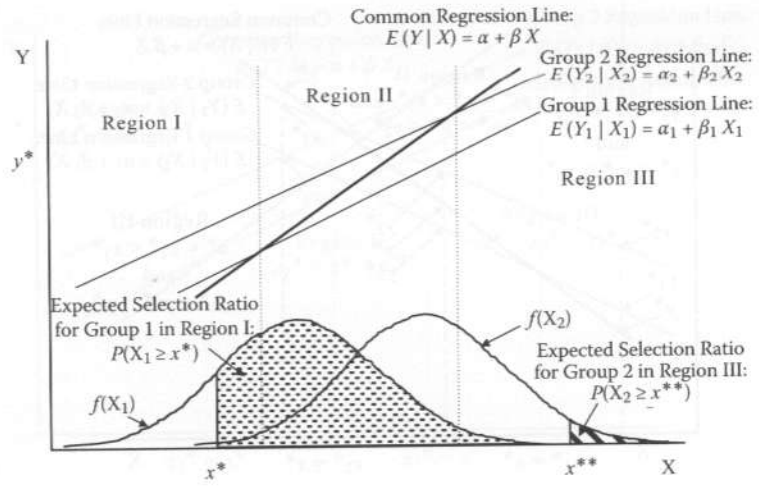

FIGURE 15.5
Expected selection ratios by test score region for intercept-based test bias.

ratios in the other two regions. For example, the largest shaded area in Figure 15.5 coincides with the expected selection ratio for Group 1 at the selection cutoff x^* in Region I. As depicted, that shaded area is about two thirds of the total area under the distribution of test scores for minority Group 1, $f(X_1)$; thus, at cutoff x^* in Region I a large percentage of applicants from Group 1 is expected to be selected. Also at x^* in Region I, note that virtually all candidates from the majority Group 2 are expected to be selected because the area under $f(X_2)$ to the right of x^* (an area that is not shaded in Figure 15.5) captures almost all of the Group 2 test score probability mass. Figure 15.5 can be used to visualize clearly what happens to expected selection ratios when selection cutoffs are increased: The expected selection ratios for both groups become increasingly small with larger cutoffs. See, for example, the smaller shaded area in Figure 15.5 depicting the expected selection ratio for Group 2 in Region III at x^{**}. Thus, large percentages of applicants are expected to be selected in Region I and smaller percentages in Region III.

Taken together, Figures 15.4 and 15.5 illustrate the kinds of trade-offs that decision makers face when using selection systems as if they were unbiased in the presence of actual test bias. Considering performance differentials only, Region I is desirable because both groups are expected to exceed performance expectations (Figure 15.4). However, this would mean that the expected selection ratios are very large (i.e., large proportions of applicants are expected to be selected from each group (Figure 15.5), which may seriously compromise the economic utility of using the test as is usually conceptualized in terms of individual performance in industrial and

organizational psychology (Cascio & Aguinis, 2005, Chapter 3). Conversely, minimizing expected selection ratios to enhance test utility, which occurs when cuts fall within Region III, would lead to the most disadvantageous results in terms of expected performance.

Now, let us discuss implications of lowering cut scores in terms of bias-based selection errors. There are two types of bias-based errors that can occur: (a) expected false positives (i.e., individuals selected do not meet standards) and (b) expected false negatives (i.e., individuals not selected who could have met the standards). We turn first to bias-based expected false positives. Expected false positives that arise from test bias occur whenever, at any given performance level, y^*, a group-specific selection cutoff, exceeds the common line selection cutoff (Aguinis & Smith, 2007). Careful scrutiny of Figure 15.3 reveals that there will be no expected false positives in Region I because, everywhere in this region, the common line cutoff x^* exceeds the group-specific cutoffs x_1^* and x_2^*. In Region II, there are expected false positives for Group 1 only. Both groups will have expected false positives in Region III.

We can ascertain the magnitude of expected false positives by using probability calculations analogous to those applied to expected selection ratios. Consider Figure 15.6 and suppose, for example, that the desired performance level is y^*. At y^*, all individuals with test scores exceeding x^* are under consideration for employment. However, over the range of test scores x^* and x_1^*, individuals from Group 1 will actually perform worse than the expected performance level y^* because the values for Y over this range are lower than y^* along the Group 1 regression line. These are

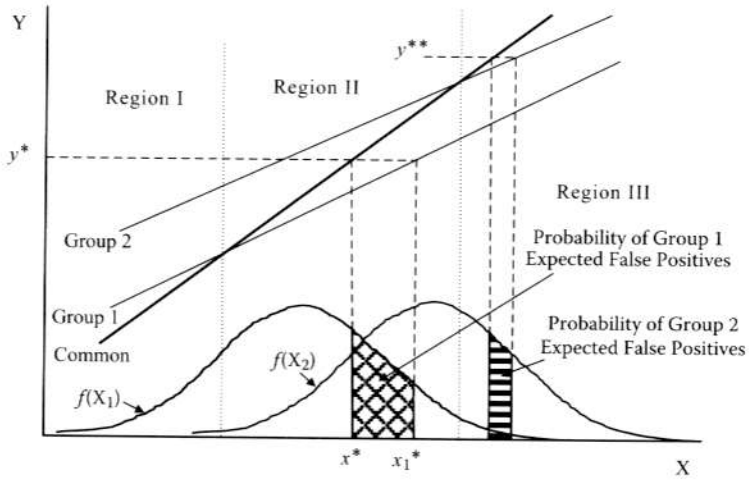

FIGURE 15.6
Expected false positives by test score region for intercept-based test bias.

expected false positives. The *probability* of expected false positives will be the area under the Group 1 test score distribution $f(X_1)$ between x^* and x_1^* as shown in Figure 15.6. Probabilities of expected false positives for Group 1 in Region II will generally be larger than those in Region III because Region III coincides with smaller probability mass regions (i.e., the tails) of $f(X_1)$. Figure 15.6 also shows how to identify probabilities of expected false positives for Group 2 in Region III, where a different performance level y^{**} exceeds the performance level predicted by the Group 2 regression line over the relevant range of test scores.

Bias-based expected false negatives occur whenever, for a given performance level, the common line cutoff exceeds a group-specific cutoff (Aguinis & Smith, 2007). Referring to Figure 15.3, we see that bias-based expected false negatives will not occur in Region III. Region II will have expected false negatives but for Group 2 only. Both groups will have expected false negatives in Region I.

Now, please refer to Figure 15.7 to consider probabilities of expected false negatives. At performance level y^*, only those applicants whose test scores exceed x^* are under consideration; those with test scores less than x^* are not. Note, however, that over the range x_2^* to x^*, performance levels at the Group 2 regression line exceed y^*; in other words, Group 2 individuals in this range exceed the expected performance level but are not being considered for employment. This is an expected false negative. Probabilities of expected false negatives are areas under group-specific test score distributions, as shown in Figure 15.7. Although Group 2 will have expected false positives in Regions I and II, they will typically be larger in Region II, where there is more probability mass.

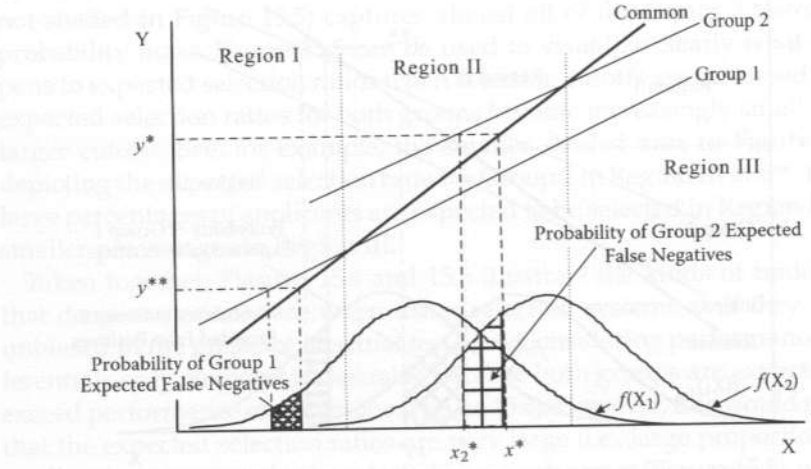

FIGURE 15.7
Expected false negatives by test score region for intercept-based test bias.

Finally, the expected AI ratio will be more severe at high selection cutoffs than at low ones for the scenarios such as those in Figures 15.3 through 15.7, in which test bias is characterized by intercept differences by group. Therefore, Region I is the most desirable, and Region III the least, as regards expected AI.

To summarize our discussion thus far, Table 15.1 shows what happens when intercept-based test bias is taken into account when cut scores are changed in an attempt to mitigate AI. This table makes the various trade-offs explicit and demonstrates that a decision to vary cut scores to address AI is more complex than has been discussed thus far in the literature. For example, if cut scores fall within Region III, the test will be highly selective

TABLE 15.1

Summary of Trade-Offs Among Expected Adverse Impact, Expected Selection Ratios, and Expected False Positives and False Negatives by Test Score Region in the Presence of Intercept-Based Test Bias

	Region I: Low selection cut scores	Region II: Moderate selection cut scores	Region III: High selection cut scores
Expected adverse impact (EAI)	*Desirable*: Large numerical values for EAI (i.e., more likely to meet the 80% heuristic)	*Moderate*	*Undesirable*: Small numerical values for EAI (i.e., more severe adverse impact)
Expected selection ratios	*Undesirable*: Larger (i.e., minimizes test utility)	*Moderate*	*Desirable*: Smaller (i.e., maximizes test utility)
Performance differentials	*Desirable*: Both groups would perform better than expected	*Mixed*: Group 1 would perform worse than expected, and Group 2 would perform better than expected	*Undesirable*: Both groups would perform worse than expected
Expected false negatives	*Undesirable* but not as severe as Region II: Both groups would have expected false negatives but tend to be small	*Mixed*: Group 2 only—can be large; *undesirable* if the primary goal is to minimize expected false negatives	*Desirable*: No expected false negatives
Expected false positives	*Desirable*: No expected false positives	*Mixed*: Group 1 only—can be large; *undesirable* if the primary goal is to minimize expected false positives	*Undesirable*, but not as severe as Region II: Both groups would have expected false positives, but tend to be small

(in the sense that selection cutoffs are large and expected selection ratios are low), and test utility will be maximized. Also on the positive side, there will be no bias-based expected false-negative errors, which would be a highly desirable outcome in a tight labor market (i.e., all applicants who are likely to succeed on the job are given a job offer). However, expected AI will be severe (likely violating the 80% heuristic) and observed performance will be worse than anticipated for both groups. In addition, there will be false positives (yet small in magnitude) for both groups.

What happens if we are faced with a Region III situation and decided to lower the cut score to reach a more acceptable level of AI? If we go from Region III to Region II, there would be expected false negatives for Group 2 (possibly large in magnitude) as well as expected false-positive errors for Group 1 (also possibly large). If AI is still not acceptable, we could decide to lower the cut scores even more and move into Region I. If this happened, Table 15.1 shows that the test would decrease its selectivity (and utility), perhaps to a level that is just unacceptable (i.e., almost all applicants would have to be selected), and there would be expected false negatives in both groups.

In closing, we have known for some time that higher cut scores are associated with more severe AI and greater test utility, whereas lower cut scores are associated with less-severe AI and less test utility (Aguinis, 2004b). Our discussion shows that the relationship between cut scores and AI is more complex, and there are several additional unanticipated consequences of changing cut scores to yield a more acceptable AI ratio. When test bias exists (even if it is small), changing cut scores leads to important consequences in terms of expected employee performance as well as expected selection errors (both false negatives and false positives) that have not been considered thus far.

To this point, we intentionally limited our discussion to the use of graphs to illustrate our points. Next, we offer two numerical cases to demonstrate the complexity of the cut score–AI relationship when test bias exists. By changing cut scores to mitigate AI, there can be unanticipated outcomes that are beneficial in terms of selection decision making (i.e., better performance than anticipated), but in other cases the unanticipated outcomes can be quite negative (i.e., larger expected false positives and negatives than anticipated).

Case 1: Intercept-Based Differences

In this first numerical example, we use the same parameters from Scenario B in Aguinis and Smith (2007). Specifically, in this situation the minority

TABLE 15.2

Summary of Trade-Offs Among Expected Adverse Impact, Expected Selection Ratios, and Expected False Positives and False Negatives by Test Score Region in the Presence of Intercept-Based Test Bias (Case 1)

	Region I: Low selection cut scores[a]	Region II: Moderate selection cutoffs	Region III: High selection cutoffs
Expected adverse impact (EAI)	N/A	Ranges from 100% at cutoff (x^*) < 54 to 17% at x^* = 117 EAI = 80% at x^* = 87	EAI is 17% to 4%
Expected selection ratios	N/A	*Group 1* ranges from 100% at x^* = 23 to 0.7% at x^* = 117.4 *Group 2* ranges from 100% at x^* = 23 to 4% at x^* = 117.4	*Group 1*: 0.7% or less *Group 2*: 4% or less
Performance differentials	N/A	Negligible; within ± 0.4 points of expected performance for both groups	*Group 1*: Underperforms by as much as 0.5 points *Group 2*: Negligible
Expected false negatives	N/A	*Group 2* ranges from zero to 6% (the latter at x^* = 96)	
Expected false positives	N/A	*Group 1* ranges from zero to 22% (the latter at x^* = 91)	Negligible; 0.6% or less for Group 1 and 0.1% or less for Group 2

Note: N/A, not applicable.

[a] Region I is out of the applicable range for this particular case.

group (i.e., Group 1) comprises 20% of the total number of applicants, has a mean score on the test of $\mu_{X1} = 92.8$, and mean performance score of $\mu_{Y1} = 2.75$ (on a 5-point scale of supervisory ratings). For the majority group (i.e., Group 2), $\mu_{X2} = 100$ and $\mu_{Y2} = 3.5$. Also, $\sigma_{X1} = \sigma_{X2} = 10$, $\sigma_{Y1} = \sigma_{Y2} = 1$, and $\rho_1 = \rho_2 = 0.5$, which, as noted by Aguinis and Smith (2007) is consistent with evidence generated by several meta-analytic reviews. Also, when the entire population is considered without breaking it down into groups, $\mu_X = 98.56$, $\sigma_X = 10.41$, $\mu_Y = 3.35$, $\sigma_Y = 1.04$, and $\rho = 0.54$.

We used the Aguinis and Smith (2007) calculator available online at http://mypage.iu.edu/~haguinis/selection/, which presumes bivariate normality of test scores and performance, to generate the values shown in Table 15.2 for each of the three relevant regions. Sample-based statistics can be used in lieu of population parameters in obtaining numerical results for actual selection situations. For the purposes of discussion, we set the lower bounds for Region I at performance level Y = –1.25 and test score X = 52.8 and the upper bounds for Region III at Y = 7.5 and X = 140.

These values are four standard deviations beyond the closest group-specific means. For this particular case, the transition from Region I to Region II occurs at X = 23.4 and Y = –0.7. Therefore, for all practical purposes of users of a test such as this one, Region I will never be encountered as it is beyond the relevant range of test scores.

Using the Web-based calculator to obtain precise numerical results based on realistic data showed that the decision to lower cut scores to mitigate AI leads to several unanticipated outcomes. As expected, using a cut score in Region III, the one with the highest degree of selectivity, leads to severe expected AI (i.e., around 17% or smaller), which obviously violates the 80% heuristic. So, selection decision makers would consider lowering the cut scores to Region II. In this region, expected AI may now fall between the 80% and 100%, which is an acceptable range. However, selectivity (and test utility) is lowered. One set of surprising results relate to performance differentials because there would be an unanticipated observed mean performance decrease of up to 0.4 points (on a 5-point scale of supervisory ratings) for the minority group and an unanticipated observed increase of up to 0.4 points for the majority group. Expected false negatives could be as high as 6% for the majority group. In terms of expected false positives, the minority group could reach as much as 22%, a potentially substantial number of workers who will not meet performance expectations. In short, for this realistic case, lowering the cut score would lead to the benefit of reaching an acceptable level of expected AI and would need to be weighed against the cost of a decrease in selectivity, a decrease in performance for the minority group (albeit small), and an increase in expected false positives for the minority group. Using the online calculator allows decision makers to obtain precise numerical results that make the trade-offs involved explicit. Consequently, the decision to lower the cut score can be made within a broader context of outcomes beyond AI.

Case 2: Intercept- and Slope-Based Differences

In this second numerical example, we use parameter values that are similar to those in Case 1, but we changed them slightly so that differences across groups are based on both intercepts and slopes. In this scenario in which the group-based regression lines are not parallel, $\mu_{X2} = 100$, $\mu_{X1} = 85$, $\sigma_{X1} = \sigma_{X2} = 20$, $\sigma_{Y1} = 1.2$, $\sigma_{Y2} = 0.8$, $\rho_1 = 0.58$, $\rho_2 = 0.49$, $\mu_{Y2} = 5$, $\mu_{Y1} = 4$, $\mu_X = 92.5$, $\sigma_X = 21.36$, $\mu_Y = 4.5$, $\sigma_Y = 1.1358$, $\rho = 0.603$, and half of the population is in Group 1 (the other half is in Group 2). Again, these parameter values are quite realistic (cf. Hunter, Schmidt, & Hunter, 1979). The group-specific

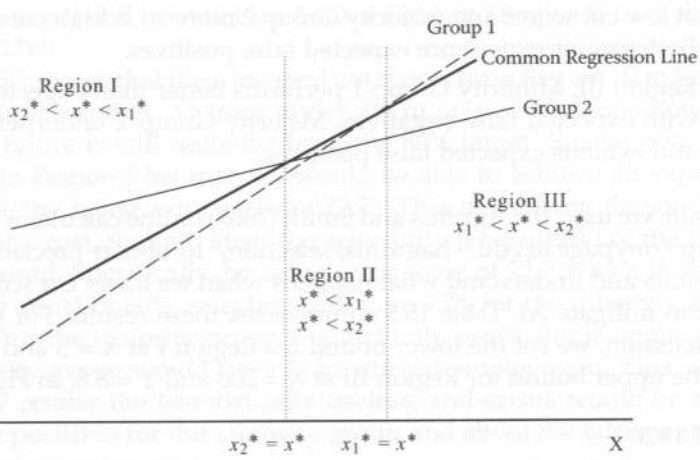

FIGURE 15.8
The three regions of test scores in the presence of intercept- and slope-based test bias (Case 2).

and common regression lines that result from these parameter values are displayed in Figure 15.8.

As we did throughout, Region I is defined as including the low cut scores, Region III includes the high cut scores, and Region II includes the intermediate range. However, given the configuration of the lines shown in Figure 15.8, the boundary between Regions I and II is now the intersection of the Group 2 and common regression lines, and the boundary separating Region II from Region III is now the intersection of the Group 1 and common regression lines. And, because the group-specific regression lines are no longer parallel, they intersect in Region II (as displayed in Figure 15.8).

Like Case 1, expected selection ratios are largest in Region I and decline with increasing cut scores. Also like Case 1, expected AI ratios are acceptable in Region I but become smaller (i.e., more severe expected AI) with increasing cut scores, so that expected AI is most severe in Region III. However, that is where the similarity between Case 1 and Case 2 ends. Specifically, as depicted Figure 15.8, we observe the following outcomes by region:

- Region I: The majority Group 2 performs better than expectations and includes expected false negatives. The minority Group 1 underperforms relative to expectations and exhibits expected false positives.
- Region II: Both the minority and majority groups are expected to underperform relative to expectations, minority Group 1 more so

at low cut scores and majority Group 2 more so at high cut scores. Both groups experience expected false positives.

- Region III: Minority Group 1 performs better than expected, but with expected false negatives. Majority Group 2 underperforms and exhibits expected false positives.

Again, we used the Aguinis and Smith (2007) online calculator available at http://mypage.iu.edu/~haguinis/selection/ to obtain precise numerical results and understand what happens when we lower cut scores in an effort to mitigate AI. Table 15.3 summarizes these results. For purposes of discussion, we set the lower bound for Region I at X = 5 and Y = −0.8 and the upper bound for Region III at X = 200 and Y = 8.8. In Figure 15.3,

TABLE 15.3

Summary of Trade-Offs Among Expected Adverse Impact, Expected Selection Ratios, and Expected False Positives and False Negatives by Test Score Region in the Presence of Intercept- and Slope-Based Test Bias (Case 2)

	Region I: Low cut scores	Region II: Moderate cut scores	Region III: High cut scores
Expected adverse Impact (EAI)	Ranges from 100% at $x^* < 7.3$ to 25% at $x^* = 120.8$ EAI = 80% at $x^* = 72.7$	25% to 3%	Under 3%
Expected selection ratios	*Group 1:* Ranges from 100% at x^*, 7.2 to 4% at $x^* = 120.8$ *Group 2:* Ranges from 100% at $x^* < 22.1$ to 15% at $x^* = 120.8$	*Group 1*: 4% or less *Group 2*: 15% or less	Virtually no one is selected in Region III; Region III is not relevant for this scenario
Performance differentials	*Group 1:* Underperforms by 0.5 points at $x^* = 5$ and by 0.2 points at $x^* = 120.8$ *Group 2*: Performs better than expected by up to 1.4 points at $x^* = 5$	*Group 1:* Underperforms by up to 0.2 points at $x^* = 120.89$ *Group 2:* Underperforms by up to 0.7 points at $x^* = 170.9$	
Expected false negatives	*Group 2*: Up to 26% at $x^* = 96$		*Group 1*: Negligible
Expected false positives	*Group 1*: Up to 16% at $x^* = 77$	Negligible: No more than 1.5% for Group 1 and 3% for Group 2	*Group 2*: Negligible

Region I is separated from Region 2 at X = 120.8 and Region II from Region III at X = 179.9.

Table 15.3 shows that if we lowered cut scores from Region III to Region II, we would improve AI from under 3% to up to 25%. Note, however, that this figure is still quite far from the 80% target. So, we may wish to move to Region I because we would be able to achieve an expected AI of 80% (by using a cut score of 72.7). This move from Region III to I comes at a cost, though. First, the selectivity (and utility) of the test is compromised. Specifically, by using a cut score of 72.7, which is associated with an AI of 80%, selection ratios are 73% for the minority group and 91% for the majority group. So, virtually every single applicant in the majority group would have to be offered employment. This would obviously render the test virtually useless, and errors would be about 15% false positives for the minority group and about 8% false negatives for the majority group.

Discussion

This chapter's main contribution is to demonstrate the complex issues involved in changing cut scores in an attempt to mitigate AI in the presence of test bias. Specifically, depending on the degree and form of test bias, lowering cut scores can help mitigate AI. However, this lowering of cut scores can also degrade the selectivity of a test, decrease a test's economic utility, and increase bias-based false positives and false-negative errors. Also depending on the situation in hand, lowering cut scores may actually lead to beneficial outcomes such as a decrease in bias-based false positives or false negatives. The Aguinis and Smith (2007) decision-making framework and online calculator can be used to understand what are the expected outcomes of a particular decision (i.e., decrease the cut score by a given amount given a particular situation, specific mean test and criterion scores for each of the groups, group-based validity coefficients, and so forth). Next, we discuss some implications for theory and research as well as practice.

Implications for Theory and Research

The scholarly literature relating cut scores and AI has thus far focused on the implications of lowering cut scores in terms of a system's selectivity and test utility. Our chapter offers an expanded and more comprehensive view of the cut score–AI relationship that includes a consideration of the presence of test bias. It would be difficult to argue that regression lines

across groups are always precisely identical. Likely, albeit small in some cases, differences across lines exist. The fact that such differences are sometimes reported (i.e., Hunter & Schmidt, 1976; Reilly, 1973; Rotundo & Sackett, 1999) in spite of the lack of statistical power for the moderating effect suggests that test bias may be more pervasive than thought (Aguinis et al., 2005). Thus, there is a need for further theory work, as well as empirical research, on the reasons why test bias exists, how to detect it, and how to mitigate it. Some efforts in this regard are quite promising (e.g., Cronshaw, Hamilton, Onyura, & Winston, 2006), but much work remains to be done.

A second implication for theory and research is that understanding where expected selection errors will occur (i.e., under which region) and the severity of such errors (i.e., percentages of false positives and negatives in each group) is not a simple process. Rather, such outcomes are understood by engaging in an inductive and interactive process in which a researcher enters values in a sort of trial-and-error fashion in the online calculator to obtain results for each scenario. Although Tables 15.2–15.3 include summary information regarding the trade-offs involved by region in two typical cases, these numerical values change based on the degree of bias that may be present. Future research could investigate thresholds for test bias that may lead to undesirable results. For example, given intercept-based test bias, how different can the regression lines be until there is a noticeable impact on, for example, expected performance for the minority group? Future research can address similar questions regarding a maximum test bias threshold that would allow for acceptable selection outcomes (e.g., false positives and false negatives).

Implications for Practice

One important implication for practice is that test bias should no longer be excluded from selection decision making in organizational staffing and high-stakes testing. Given the availability of the Aguinis and Smith (2007) online calculator, there is no reason not to use it to anticipate the impact of lowering cut scores on such crucial selection outcomes as AI, differences between anticipated and observed performance in those hired, and selection errors, including bias-based false positives and false negatives. If test bias does not have an important effect on these outcomes given a specific situation, then the online calculator will show that. On the other hand, if bias is present (even if it is small in magnitude), the online calculator will consider its effects when computing the anticipated outcomes. Practitioners have the professional mandate to make the best possible decision in terms of selection, particularly when high-stakes testing is involved. Using the online calculator allows for the consideration of possible test bias and its effects on important selection outcomes explicitly. In

many cases, and depending on the particular situation, using the online calculator may show that the cure (i.e., lowering cut scores) may actually be worse than the disease (i.e., AI). In fact, our analyses showed that there is no cut score region in which some kind of unpleasant outcome does not occur when test bias is present.

Another implication of our analyses is that, given typical test characteristics that we discussed, it is virtually impossible to use a GMA test and reach the 80% AI heuristic without hiring such a large proportion of applicants that the utility of the test is compromised. Stated differently, how many GMA tests can be used with cut scores in Region III and yet lead to minimum, or even acceptable, AI? This is a challenge for practitioners but obviously is linked to a need for further research to solve this problem.

In closing, the possibility of test bias must be taken into account before deciding to lower a cut score to mitigate AI. The Aguinis and Smith (2007) online calculator allows researchers and practitioners to consider specific numerical characteristics of a testing situation and compute anticipated selection outcomes, including AI, differences in expected versus observed performance for those who will be hired, and false-positive and false-negative selection errors. Obtaining these numbers and considering them explicitly before a test is put to use will help improve organizational staffing and high-stakes selection decisions.

Acknowledgment

This research was conducted, in part, while Herman Aguinis held the Mehalchin Term Professorship in Management at the University of Colorado at Denver and visiting appointments at the University of Salamanca (Spain) and University of Puerto Rico.

References

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

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

Aguinis, H. (Ed.). (2004b). *Test-score banding in human resource selection: Legal, technical, and societal issues.* Westport, CT: Praeger.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (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, R. J., & Pierce, C. A. (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., Cortina, J. M., & Goldberg, E. (1998). A new procedure for computing equivalence bands in personnel selection. *Human Performance, 11,* 351–365.

Aguinis, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology, 60,* 165–199.

Aguinis, H., & Stone-Romero, E. F. (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, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: 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, P. L. (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 J. Martocchio (Ed.), *Research in personnel and human resources management* (Vol. 19, pp. 177–197). New York: Elsevier.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

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

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

Cronshaw, S. F., Hamilton, L. K., Onyura, B. R., & Winston, A. S. (2006). The case for non-biased intelligence testing against Black Africans has not been made: A comment on Rushton, Skuy, and Bons (2004). *International Journal of Selection and Assessment, 14,* 278–287.

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

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review an analysis. *Psychological Bulletin, 86,* 721–735.

Kehoe, J. (2009). Cut scores and adverse impact. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 293–328). New York: Routledge.

Murphy, K. R. (2004). Conflicting values and interests in banding research and practice. In H. Aguinis (Ed.), *Test-score banding in human resource selection: Legal, technical, and societal issues* (pp. 175–192). Westport, CT: Praeger.

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

Roth, P. L., Bobko, P., & Switzer, F. S. (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology, 91*, 507–522.

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

Schmidt, F. L., & Hunter, J. E. (2004). SED banding as a test of scientific values in I/O psychology. In H. Aguinis (Ed.), *Test-score banding in human resource selection: Legal, technical, and societal issues* (pp. 151–174). Westport, CT: Praeger.

Uniform Guidelines on Employee Selection Procedures. 43 Fed. Reg. 38290–38315 (1978).

Zedeck, S. (2009). Adverse impact: History and evolution. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 3–28). New York: Routledge.