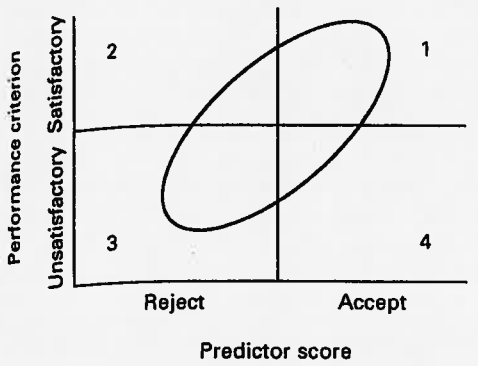


FIGURE 8-1 Positive validity.

Figure 8-1 shows that the relationship is positive and people with high (low) predictor scores also tend to have high (low) criterion scores. In investigating differential validity for groups (e.g., ethnic minority and ethnic nonminority), if the joint distribution of predictor and criterion scores is similar throughout the scatterplot in each group, as in Figure 8-1, no problem exists, and use of the predictor can be continued. On the other hand, if the joint distribution of predictor and criterion scores is similar for each group, but circular, as in Figure 8-2, there is also no differential validity, but the predictor is useless because it supplies no information of a predictive nature. So there is no point in investigating differential validity in the absence of an overall pattern of predictor–criterion scores that allows for the prediction of relevant criteria.

Differential Validity and Adverse Impact

An important consideration in assessing differential validity is whether the test in question produces adverse impact. The *Uniform Guidelines* (1978) state that a "selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or 80 percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact" (p. 123). In other words, adverse impact means that members of one group are selected at substantially greater rates than members of another group. To understand whether this is the case, one compares selection ratios across the groups under consideration. For example, assume that the applicant pool consists of 300 ethnic minorities and 500 nonminorities. Further, assume that 30 minorities are hired, for a selection ratio of $SR_1 = 30/300 = 10$, and that 100 nonminorities are hired, for a selection ratio of $SR_2 = 100/500 = 20$. The adverse impact ratio is $SR_1/SR_2 = .50$, which is substantially smaller than the recommended .80 ratio. Let's consider various

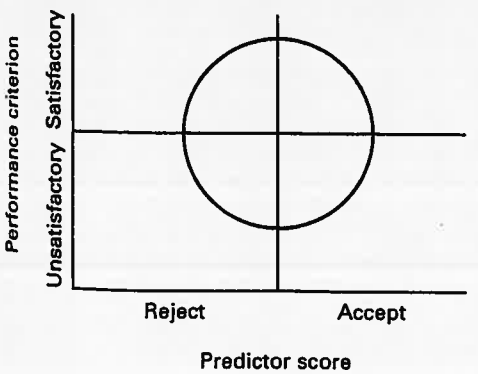


FIGURE 8-2 Zero validity.

170

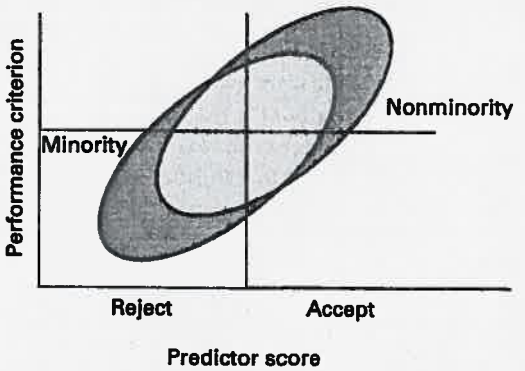


FIGURE 8-3 Valid predictor with adverse impact.

scenarios relating differential validity with adverse impact. The ideas for many of the following diagrams are derived from Barrett (1967) and represent various combinations of the concepts illustrated in Figures 8-1 and 8-2.

Figure 8-3 is an example of a differential predictor–criterion relationship that is legal and appropriate. In this figure, validity for the minority and nonminority groups is equivalent, but the minority group scores lower on the predictor and does poorer on the job (of course, the situation could be reversed). In this instance, the very same factors that depress test scores may also serve to depress job performance scores. Thus, adverse impact is defensible in this case, since minorities do poorer on what the organization considers a relevant and important measure of job success. On the other hand, government regulatory agencies probably would want evidence that the criterion was relevant, important, and not itself subject to bias. Moreover, alternative criteria that result in less adverse impact would have to be considered, along with the possibility that some third factor (e.g., length of service) did not cause the observed difference in job performance (Byham & Spitzer, 1971).

An additional possibility, shown in Figure 8-4, is a predictor that is valid for the combined group, but invalid for each group separately. In fact, there are several situations where the validity coefficient is zero or near zero for each of the groups, but the validity coefficient in both groups combined is moderate or even large (Ree, Carretta, & Earles, 1999). In most cases where no validity exists for either group individually, errors in selection would result from using the predictor without validation or from failing to test for differential validity in the first place. The predictor in this case becomes solely a crude measure of the grouping variable (e.g., ethnicity) (Bartlett & O'Leary, 1969). This is the most clear-cut case of using selection measures to discriminate in terms of race, sex, or any other unlawful basis. Moreover, it is unethical to use a selection device that has not been validated (see Appendix A).

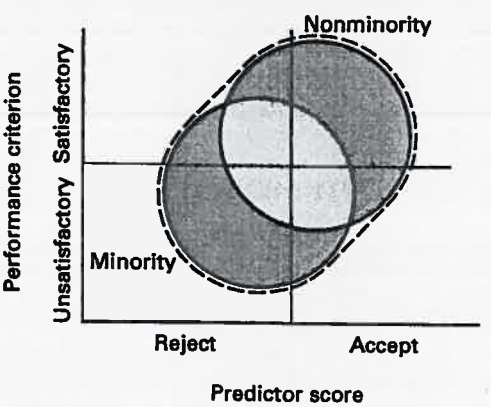


FIGURE 8-4 Valid predictor for entire group; invalid for each group separately.

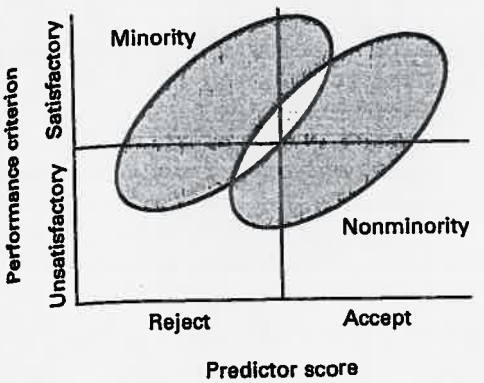


FIGURE 8-5 Equal validity, unequal predictor means.

It also is possible to demonstrate equal validity in the two groups combined with unequal predictor means or criterion means and the presence or absence of adverse impact. These situations, presented in Figures 8-5 and 8-6, highlight the need to examine differential prediction, as well as differential validity.

In Figure 8-5, members of the minority group would not be as likely to be selected, even though the probability of success on the job for the two groups is essentially equal. Under these conditions, an alterative strategy is to use separate cut scores in each group based on predictor performance, while the expectancy of job performance success remains equal. Thus, a Hispanic candidate with a score of 65 on an interview may have a 75 percent chance of success on the job. A white candidate with a score of 75 might have the same 75 percent probability of success on the job. Although this situation might appear disturbing initially, remember that the predictor (e.g., a selection interview) is being used simply as a vehicle to forecast the likelihood of successful job performance. The primary focus is on job performance rather than on predictor performance. Even though interview scores may mean different things for different groups, as long as the expectancy of success on the job is equal for the two (or more) groups, the use of separate cut scores is justified. Indeed, the reporting of an expectancy score for each candidate is one recommendation made by a National Academy of Sciences panel with respect to the interpretation of scores on the General Aptitude Test Battery (Hartigan & Wigdor, 1989). A legal caveat exists, however. In the United States, it is illegal to use different selection rules for identifiable groups in some contexts (Sackett & Wilk, 1994).

Figure 8-6 depicts a situation where, although there is no noticeable difference in predictor scores, nonminority group members tend to perform better on the job than minority group members (or vice versa). If predictions were based on the combined sample, the result would be a systematic

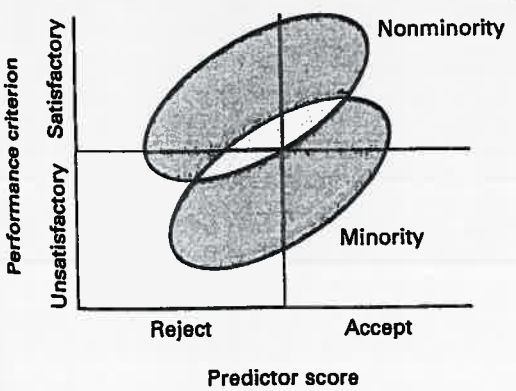


FIGURE 8-6 Equal validity, unequal criterion means.

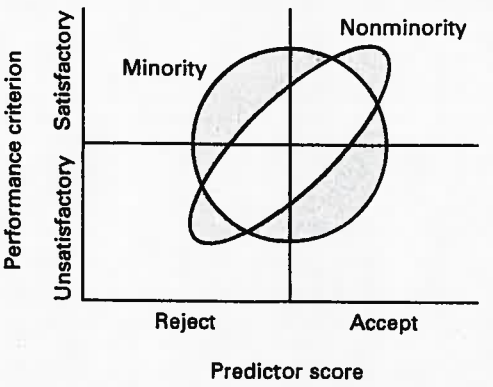
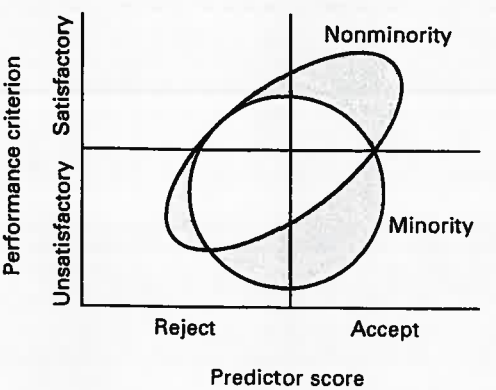


FIGURE 8-7 Equal predictor means, but validity only for the nonminority group.

underprediction for nonminorities and a systematic overprediction for minorities, although there is no adverse impact. Thus, in this situation, the failure to use different selection rules (which would yield more accurate prediction for both groups) may put minority persons in jobs where their probability of success is low and where their resulting performance only provides additional evidence that helps maintain prejudice (Bartlett & O'Leary, 1969). The nonminority individuals also suffer. If a test is used as a placement device, for example, since nonminority performance is systematically underpredicted, these individuals may well be placed in jobs that do not make the fullest use of their talents.

In Figure 8-7, no differences between the groups exist either on predictor or on criterion scores; yet the predictor has validity only for the nonminority group. Hence, if legally admissible, the selection measure should be used only with the nonminority group, since the job performance of minorities cannot be predicted accurately. If the measure were used to select both minority and nonminority applicants, no adverse impact would be found, since approximately the same proportion of applicants would be hired from each group. However, more nonminority members would succeed on the job, thereby reinforcing past stereotypes about minority groups and hindering future attempts at equal employment opportunity (EEO).

In our final example (see Figure 8-8), the two groups differ in mean criterion performance as well as in validity. The predictor might be used to select nonminority applicants, but should not be used to select minority applicants. Moreover, the cut score or decision rule used to select nonminority applicants must be derived solely from the nonminority group, *not* from the combined group. If the minority group (for whom the predictor is not valid) is included, overall validity will be lowered, as will the overall mean criterion score. Predictions will be less accurate because the standard error of estimate will be inflated. As in the previous example, the organization should use the selection measure only for the nonminority group (taking



SLOPES NOT PARALLEL

FIGURE 8-8 Unequal criterion means and validity, only for the nonminority group.

into account the caveat above about legal standards) while continuing to search for a predictor that accurately forecasts minority job performance. Recall that the Civil Rights Act of 1991 (see Chapter 2) makes it unlawful to use different cutoff scores on the basis of race, color, religion, sex, or national origin. However, an employer may make test-score adjustments as a consequence of a court-ordered affirmative action plan or where a court approves a conciliation agreement.

In summary, numerous possibilities exist when heterogeneous groups are combined in making predictions. When differential validity exists, the use of a single regression line, cut score, or decision rule can lead to serious errors in prediction. While one legitimately may question the use of race or gender as a variable in selection, the problem is really one of distinguishing between performance on the selection measure and performance on the job (Guion, 1965). If the basis for hiring is expected job performance and if different selection rules are used to improve the prediction of expected job performance rather than to discriminate on the basis of race, gender, and so on, then this procedure appears both legal and appropriate. Nevertheless, the implementation of differential systems is difficult in practice because the fairness of any procedure that uses different standards for different groups is likely to be viewed with suspicion ("More," 1989).

Differential Validity: The Evidence

Let us be clear at the outset that evidence of differential validity provides information only on whether a selection device should be used to make comparisons *within* groups. Evidence of unfair discrimination *between* subgroups cannot be inferred from differences in validity alone; mean job performance also must be considered. In other words, a selection procedure may be fair and yet predict performance inaccurately, or it may discriminate unfairly and yet predict performance within a given subgroup with appreciable accuracy (Kirkpatrick, Ewen, Barrett, & Katzell, 1968).

In discussing differential validity, we must first specify the criteria under which differential validity can be said to exist at all. Thus, Boehm (1972) distinguished between differential and single-group validity. Differential validity exists when (1) there is a significant difference between the validity coefficients obtained for two subgroups (e.g., ethnicity or gender) and (2) the correlations found in one or both of these groups are significantly different from zero. Related to, but different from, differential validity is single-group validity, in which a given predictor exhibits validity significantly different from zero for one group only, and there is no significant difference between the two validity coefficients.

Humphreys (1973) has pointed out that single-group validity is not equivalent to differential validity, nor can it be viewed as a means of assessing differential validity. The logic underlying this distinction is clear: To determine whether two correlations differ from each other, they must be compared directly with each other. In addition, a serious statistical flaw in the single-group validity paradigm is that the sample size is typically smaller for the minority group, which reduces the chances that a statistically significant validity coefficient will be found in this group. Thus, the appropriate statistical test is a test of the null hypothesis of zero difference between the sample-based estimates of the population validity coefficients. However, statistical power is low for such a test, and this makes a Type II error (i.e., not rejecting the null hypothesis when it is false) more likely. Therefore, the researcher who unwisely does not compute statistical power and plans research accordingly is likely to err on the side of *too few* differences. For example, if the true validities in the populations to be compared are .50 and .30, but both are attenuated by a criterion with a reliability of .7, then even without any range restriction at all, one must have 528 persons *in each group* to yield a 90 percent chance of detecting the existing differential validity at alpha = .05 (for more on this, see Trattner & O'Leary, 1980).

The sample sizes typically used in any one study are, therefore, inadequate to provide a meaningful test of the differential validity hypothesis. However, higher statistical power is possible if validity coefficients are cumulated across studies, which can be done using meta-analysis (as discussed in Chapter 7). The bulk of the evidence suggests that statistically significant differential

validity is the exception rather than the rule (Schmidt, 1988; Schmidt & Hunter, 1981; Wigdor & Garner, 1982). In a comprehensive review and analysis of 866 black–white employment test validity pairs, Hunter, Schmidt, and Hunter (1979) concluded that findings of apparent differential validity in samples are produced by the operation of chance and a number of statistical artifacts. True differential validity probably does not exist. In addition, no support was found for the suggestion by Boehm (1972) and Bray and Moses (1972) that findings of validity differences by race are associated with the use of subjective criteria (ratings, rankings, etc.) and that validity differences seldom occur when more objective criteria are used.

Similar analyses of 1,337 pairs of validity coefficients from employment and educational tests for Hispanic Americans showed no evidence of differential validity (Schmidt, Pearlman, & Hunter, 1980). Differential validity for males and females also has been examined. Schmitt, Mellon, and Bylenga (1978) examined 6,219 pairs of validity coefficients for males and females (predominantly dealing with educational outcomes) and found that validity coefficients for females were slightly (<.05 correlation units), but significantly larger than coefficients for males. Validities for males exceeded those for females only when predictors were less cognitive in nature, such as high school experience variables. Schmitt et al. (1978) concluded: "The magnitude of the difference between male and female validities is very small and may make only trivial differences in most practical situations" (p. 150).

In summary, available research evidence indicates that the existence of differential validity in well-controlled studies is rare. Adequate controls include large enough sample sizes in each subgroup to achieve statistical power of at least .80; selection of predictors based on their logical relevance to the criterion behavior to be predicted; unbiased, relevant, and reliable criteria; and cross-validation of results.

ASSESSING DIFFERENTIAL PREDICTION AND MODERATOR VARIABLES

The possibility of predictive bias in selection procedures is a central issue in any discussion of fairness and EEO. As we noted earlier, these issues require a consideration of the equivalence of prediction systems for different groups. Analyses of possible differences in slopes or intercepts in subgroup regression lines result in more thorough investigations of predictive bias than does analysis of differential validity alone because the overall regression line determines how a test is used for prediction.

Lack of differential validity, in and of itself, does not assure lack of predictive bias. Specifically the *Standards* (AERA, APA, & NCME, 1999) note: "When empirical studies of differential prediction of a criterion for members of different groups are conducted, they should include regression equations (or an appropriate equivalent) computed separately for each group or treatment under consideration or an analysis in which the group or treatment variables are entered as moderator variables" (Standard 7.6, p. 82). In other words, when there is differential prediction based on a grouping variable such as gender or ethnicity, this grouping variable is called a **moderator**. Similarly, the 1978 *Uniform Guidelines on Employee Selection Procedures* (Ledvinka, 1979) adopt what is known as the Cleary (1968) model of test bias:

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the use of the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance. (p. 115)

In Figure 8-3, although there are two separate ellipses, one for the minority group and one for the nonminority, a single regression line may be cast for both groups. So this test would

demonstrate lack of differential prediction or predictive bias. In Figure 8-6, however, the manner in which the position of the regression line is computed clearly does make a difference. If a single regression line is cast for both groups (assuming they are equal in size), criterion scores for the nonminority group consistently will be *underpredicted*, while those of the minority group consistently will be *overpredicted*. In this situation, there is differential prediction, and the use of a single regression line is inappropriate, but it is the nonminority group that is affected adversely. While the slopes of the two regression lines are parallel, the intercepts are different. Therefore, the same predictor score has a different predictive meaning in the two groups. A third situation is presented in Figure 8-8. Here the slopes are not parallel. As we noted earlier, the predictor clearly is inappropriate for the minority group in this situation. When the regression lines are not parallel, the predicted performance scores differ for individuals with identical test scores. Under these circumstances, once it is determined where the regression lines cross, the amount of over- or underprediction depends on the position of a predictor score in its distribution.

So far, we have discussed the issue of differential prediction graphically. However, a more formal statistical procedure is available. As noted in ***Principles for the Validation and Use of Personnel Selection Procedures*** (SIOP, 2003), "testing for predictive bias involves using moderated multiple regression, where the criterion measure is regressed on the predictor score, subgroup membership, and an interaction term between the two" (p. 32). In symbols, and assuming differential prediction is tested for two groups (e.g., minority and nonminority), the moderated multiple regression (MMR) model is the following:

$$\hat{Y} = a + b_1X + b_2Z + b_3X \cdot Z \qquad \textbf{(8-1)}$$

where $\hat{Y}$ is the predicted value for the criterion Y, a is the least-squares estimate of the intercept, b_1 is the least-squares estimate of the population regression coefficient for the predictor X, b_2 is the least-squares estimate of the population regression coefficient for the moderator Z, and b_3 is the least-squares estimate of the population regression coefficient for the product term, which carries information about the moderating effect of Z (Aguinis, 2004b). The moderator Z is a categorical variable that represents the binary subgrouping variable under consideration. MMR can also be used for situations involving more than two groups (e.g., three categories based on ethnicity). To do so, it is necessary to include $k - 2$ Z variables (or code variables) in the model, where k is the number of groups being compared.

Aguinis (2004b) described the MMR procedure in detail, covering such issues as the impact of using dummy coding (e.g., minority: 1, nonminority: 0) versus other types of coding on the interpretation of results. Assuming dummy coding is used, the statistical significance of b_3, which tests the null hypothesis that $\beta_3 = 0$, indicates whether the slope of the criterion on the predictor differs across groups. The statistical significance of b_2, which tests the null hypothesis that $\beta_2 = 0$, tests the null hypothesis that groups differ regarding the intercept. Alternatively, one can test whether the addition of the product term to an equation, including the first-order effects of X and Z, only produces a statistically significant increment in the proportion of variance explained for Y (i.e., R^2).

Lautenschlager and Mendoza (1986) noted a difference between the traditional "step-up" approach, consisting of testing whether the addition of the product term improves the prediction of Y above and beyond the first-order effects of X and Z, and a "step-down" approach. The step-down approach consists of making comparisons between the following models (where all terms are as defined for Equation 8-1 above):

1: $\hat{Y} = a + b_1X$
2: $\hat{Y} = a + b_1X + b_2Z + b_3X \cdot Z$
3: $\hat{Y} = a + b_1X + b_3X \cdot Z$
4: $\hat{Y} = a + b_1X + b_3X \cdot Z$

First, one can test the overall hypothesis of differential prediction by comparing R^2s resulting from model 1 versus model 2. If there is a statistically significant difference, we then

explore whether differential prediction is due to differences in slopes, intercepts, or both. For testing differences in slopes, we compare model 4 with model 2, and, for differences in intercepts, we compare model 3 with model 2. Lautenschlager and Mendoza (1986) used data from a military training school and found that using a step-up approach led to the conclusion that there was differential prediction based on the slopes only, whereas using a step-down approach led to the conclusion that differential prediction existed based on the presence of both different slopes and different intercepts.

Differential Prediction: The Evidence

When prediction systems are compared, slope-based differences are typically not found, and intercept-based differences, if found, are such that they favor members of the minority group (i.e., overprediction of performance for members of the minority group) (Kuncel & Sackett, 2007; Rotundo & Sackett, 1999; Rushton & Jensen, 2005; Sackett & Wilk, 1994; Schmidt & Hunter, 1998; Sackett, Schmitt, Ellingson, & Kablin, 2001). Aguinis, Culpepper, and Pierce (2010) concluded that the same result has been obtained regarding selection tools used in both work and educational settings to assess a diverse set of constructs ranging from general mental abilities (GMAs) to personality and safety suitability. As they noted: "It is thus no exaggeration to assert that the conclusion that test bias generally does not exist but, when it exists, it involves intercept differences favoring minority group members and not slope differences, is an established fact in I/O psychology and related fields concerned with high-stakes testing." For example, Bartlett, Bobko, Mosier, and Hannan (1978) reported results for differential prediction based on 1,190 comparisons indicating the presence of significant slope differences in about 6 percent and significant intercept differences in about 18 percent of the comparisons. In other words, some type of differential prediction was found in about 24 percent of the tests. Most commonly, the prediction system for the nonminority group slightly overpredicted minority group performance. That is, minorities would tend to do less well on the job than their test scores predict, so there is no apparent unfairness against minority group members.

Similar results have been reported by Hartigan and Wigdor (1989). In 72 studies on the General Ability Test Battery (GATB), developed by the U.S. Department of Labor, where there were at least 50 African American and 50 nonminority employees (average sample sizes of 87 and 166, respectively), slope differences occurred less than 3 percent of the time and intercept differences about 37 percent of the time. However, use of a single prediction equation for the total group of applicants would not provide predictions that were biased against African American applicants, for using a single prediction equation slightly overpredicted performance by African Americans. In 220 tests each of the slope and intercept differences between Hispanics and nonminority group members, about 2 percent of the slope differences and about 8 percent of the intercept differences were significant (Schmidt et al., 1980). The trend in the intercept differences was for the Hispanic intercepts to be lower (i.e., overprediction of Hispanic job performance), but firm support for this conclusion was lacking.

With respect to gender differences in performance on physical ability tests, there were no significant differences in prediction systems for males and females in the prediction of performance on outside telephone-craft jobs (Reilly, Zedeck, & Tenopyr, 1979). However, considerable differences were found on both test and performance variables in the relative performances of men and women on a physical ability test for police officers (Arvey, Landon, Nutting, & Maxwell, 1992). If a common regression line was used for selection purposes, then women's job performance would be systematically overpredicted.

Differential prediction has also been examined for tests measuring constructs other than GMAs. For instance, an investigation of three personality composites from the U.S. Army's instrument to predict five dimensions of job performance across nine military jobs found that differential prediction based on sex occurred in about 30 percent of the cases (Saad & Sackett, 2002). Differential prediction was found based on the intercepts, and not the slopes. Overall,

there was overprediction of women's scores (i.e., higher intercepts for men). Thus, the result regarding the overprediction of women's performance parallels that of research investigating differential prediction by race in the GMA domain (i.e., there is an overprediction for women as there is overprediction for ethnic minorities).

Could it be that researchers find lack of differential prediction in part because the criteria themselves are biased? Rotundo and Sackett (1999) examined this issue by testing for differential prediction in the ability-performance relationship (as measured using the GATB) in samples of African American and white employees. The data allowed for between-people and within-people comparisons under two conditions: (1) when a white supervisor rated all employees, and (2) when a supervisor of the same self-reported race as each employee assigned the rating. The assumption was that, if performance data are provided by supervisors of the same ethnicity as the employees being rated, the chances that the criteria are biased are minimized or even eliminated. Analyses including 25,937 individuals yielded no evidence of predictive bias against African Americans.

In sum, the preponderance of the evidence indicates an overall lack of differential prediction based on ethnicity and gender for cognitive abilities and other types of tests (Hunter & Schmidt, 2000). When differential prediction is found, results indicate that differences lie in intercept differences and not slope differences across groups and that the intercept differences are such that the performance of women and ethnic minorities is typically overpredicted, which means that the use of test scores supposedly favors these groups.

Problems in Testing for Differential Prediction

In spite of the consistent findings, Aguinis et al. (2010) argued in favor of revival of differential prediction research because research conclusions based on work conducted over five decades on differential prediction may not be warranted. They provided analytic proof that the finding of intercept-based differences favoring minority-group members may be a statistical artifact. Also, empirical evidence gathered over the past two decades suggests that the slope-based test is typically conducted at low levels of statistical power (Aguinis, 1995, 2004b).

Low power for the slope-based test typically results from the use of small samples, but is also due to the interactive effects of various statistical and methodological artifacts such as unreliability, range restriction, and violation of the assumption that error variances are homogeneous (Aguinis & Pierce, 1998a). The net result is a reduction in the size of observed moderating effects vis-à-vis population effects (Aguinis, Beaty, Boik, & Pierce, 2005). In practical terms, low power affects fairness assessment in that one may conclude *incorrectly* that a selection procedure predicts outcomes equally well for various subgroups based on race or sex—that is, that there is no differential relationship. However, this sample-based conclusion may be incorrect. In fact, the selection procedure actually may predict outcomes differentially across subgroups. Such differential prediction may not be detected, however, because of the low statistical power inherent in test validation research.

Consider the impact of a selected set of factors known to affect the power of MMR. Take, for instance, heterogeneity of sample size across groups. In validation research, it is typically the case that the number of individuals in the minority and female groups is smaller than the number of individuals in the majority and male groups. A Monte Carlo simulation demonstrated that in differential prediction tests that included two groups there was a considerable decrease in power when the size of group 1 was .10 relative to total sample size, *regardless of total sample size* (Stone-Romero, Alliger, & Aguinis, 1994). A proportion of .30, closer to the optimum value of .50, also reduced the statistical power of MMR, but to a lesser extent. Another factor known to affect power is heterogeneity of error variance. MMR assumes that the variance in Y that remains after predicting Y from X is equal across k moderator-based subgroups (see Aguinis & Pierce, 1998a, for a review). Violating the homogeneity-of-error variance assumption has been identified as a factor that can affect the power of MMR to detect

test unfairness. In each group, the error variance is estimated by the mean square residual from the regression of Y on X:

$$\sigma_{ei}^2 = \sigma_{Y(i)}(1 - \rho_{XY(i)}^2), \tag{8-2}$$

where $\sigma_{Y(i)}$ and $\rho_{XY(i)}$ are the Y standard deviation and the X-Y correlation in each group, respectively. In the presence of a moderating effect in the population, the X-Y correlations for the two moderator-based subgroups differ, and, thus, the error terms necessarily differ.

Heterogeneous error variances can affect both Type I error (*incorrectly* concluding that the selection procedures are unfair) and statistical power. However, Alexander and DeShon (1994) showed that, when the subgroup with the larger sample size is associated with the larger error variance (i.e., the smaller X-Y correlation), statistical power is lowered markedly. Aguinis and Pierce (1998a) noted that this specific scenario, in which the subgroup with the larger n is paired with the smaller correlation coefficient, is the most typical situation in personnel selection research in a variety of organizational settings. As a follow-up study, Aguinis, Petersen, and Pierce (1999) conducted a review of articles that used MMR during 1987 and 1999 in *Academy of Management Journal*, *Journal of Applied Psychology*, and *Personnel Psychology*. Results revealed that violation of the homogeneity-of-variance assumption occurred in approximately 50 percent of the MMR tests! In an examination of error-variance heterogeneity in tests of differential prediction based on the GATB, Oswald, Saad, and Sackett (2000) concluded that enough heterogeneity was found to urge researchers investigating differential prediction to check for compliance with the assumption and consider the possibility of alternative statistical tests when the assumption is violated.

Can we adopt a meta-analytic approach to address the low-power problem of the differential prediction test? Although, in general, meta-analysis can help mitigate the low-power problem, as it has been used for testing differential validity (albeit imperfectly), conducting a meta-analysis of the differential prediction literature is virtually impossible because regression coefficients are referenced to the specific metrics of the scales used in each study. When different measures are used, it is not possible to cumulate regression coefficients across studies, even if the same construct (e.g., general cognitive abilities) is measured. This is why meta-analysts prefer to cumulate correlation coefficients, as opposed to regression coefficients, across studies (Raju, Pappas, & Williams, 1989). One situation where a meta-analysis of differential prediction tests is possible is where the same test is administered to several samples and the test developer has access to the resulting database.

Regarding the intercept-based test, Aguinis et al. (2010) conducted a Monte Carlo simulation including 3,185,000 unique combinations of a wide range of values for intercept- and slope-based test bias in the population, total sample size, proportion of minority group sample size to total sample size, predictor (i.e., preemployment test scores) and criterion (i.e., job performance) reliability, predictor range restriction, correlation between predictor scores and the dummy-coded grouping variable (e.g., ethnicity, gender), and mean difference between predictor scores across groups. Results based on 15 billion 925 million individual samples of scores suggest that intercept-based differences favoring minority group members are likely to be "found" when they do not exist. And, when they exist in the population, they are likely to be exaggerated in the samples used to assess possible test bias. The simulation results indicate that as differences in test scores between the groups increase and test-score reliability decreases, Type I error rates that indicate intercept-based differences favoring minority-group members also increase. In short, for typical conditions in preemployment testing, researchers are likely to conclude that there is intercept-based bias favoring minority group members when this is actually not true or that differences are larger than they are in actuality.

The established conclusions regarding test bias are convenient for test vendors, users, consultants, and researchers. If tests are not biased, or if they favor minority-group

members, then the fact that, on average, minority-group members score lower on GMA tests than those of the majority group is not necessarily a hindrance for the use of such tests. As noted by Kehoe (2002), "a critical part of the dilemma is that GMA-based tests are generally regarded as unbiased" (p. 104). If test bias does not exist against members of ethnic minority groups, then adverse impact against ethnic minorities is a defensible position that has formidable social consequences, and the field will continue to try to solve what seems to be an impossible dilemma between validity and adverse impact (Aguinis, 2004c; Ployhart & Holtz, 2008). A cynical approach to testing would be to perpetuate ethnic-based differences regarding GMA such that minority-group members obtain scores on average lower than majority-group members, to continue to develop tests that are less than perfectly reliable, and to assess potential test bias using the accepted Cleary (1968) regression model. This approach would make "the test 'look good' in the sense that it decreases the likelihood of observing an underprediction for the low-scoring group" (Linn & Werts, 1971, p. 3). Such a cynical approach would guarantee that slope-based differences will not be found and, if intercept-based differences are found, they will appear to favor minority-group members. In other words, there would be no charge that tests are biased against ethnic minority-group members.

In short, Aguinis et al. (2010) challenged conclusions based on 40 years of research on test bias in preemployment testing. Their results indicate that the established and accepted procedure to assess test bias is itself biased: Slope-based bias is likely to go undetected, and intercept-based bias favoring minority-group members is likely to be "found" when, in fact, it does not exist. Preemployment testing is often described as the cradle of the I/O psychology field (e.g., Landy & Conte, 2007). These results open up an important opportunity for I/O psychology researchers to revive the topic of test bias and make contributions with measurable and important implications for organizations and society (cf. Griffore, 2007; Helms, 2006).

Suggestions for Improving the Accuracy of Slope-based Differential Prediction Assessment

Fortunately, there are several remedies for the low-power problem of MMR. Table 8-1 lists several factors that lower the power of MMR, together with recommended strategies to address each of these factors. As shown in this table, there are several strategies available, but they come at a cost. Thus, HR researchers should evaluate the practicality of implementing each strategy. Luckily, there are computer programs available online that can be used to compute power before a study is conducted and that allow a researcher to investigate the pros and cons of implementing various scenarios (Aguinis, Boik, & Pierce, 2001; http://mypage.iu.edu/~haguinis/mmr/index.html). For example, one can compute the power resulting from increasing the sample size by 20 percent as compared to increasing the reliability of the predictor scores by increasing the measure's length by 30 percent. Given the cost associated with an increase in sample size vis-à-vis the improvement in predictor reliability, which of these strategies would be more cost-effective in terms of improving power? One thing is clear, however. If one waits until a validation study is finished to start thinking about statistical power for the differential prediction test, then it is probably too late. Statistical power needs to be considered long before the data are collected.

In summary, although it is reassuring to know that differential prediction does not occur often when subgroups are compared, it has been found often enough to create concern for possible predictive bias when a common regression line is used for selection. In addition, recent research has uncovered the fact that numerous statistical artifacts decrease the ability to detect differential prediction, even when it exists in the population. What's the bottom line? Carefully plan a validation study so that the differential prediction test is technically feasible and the results credible.