The Federal *Uniform Guidelines on Employee Selection Procedures* (1978)

An Update on Selected Issues

WAYNE F. CASCIO
HERMAN AGUINIS
University of Colorado at Denver

The purpose of this article is to provide an update on a selected set of issues that might be considered if and when the Uniform Guidelines on Employee Selection Procedures *is revised. The following issues that have been subject to a considerable number of research-based advances in the field of industrial and organizational psychology are discussed: (a) adverse impact, the four-fifths rule, and statistical significance testing; (b) criterion measures; (c) cutoff scores; and (d) differential prediction. In addition, implications for practice of research findings in each of these areas are discussed.*

Since the federal *Uniform Guidelines on Employee Selection Procedures* were issued in 1978, there have been a considerable number of research-based advances in the field of industrial and organizational (I&O) psychology that are relevant to various sections of the Guidelines. The purpose of this article is not to provide an exhaustive discussion of all research that is remotely relevant (e.g., cross-validation, validity generalization, alternative methods for estimating reliability, the development of the O*Net). Rather, our goal is to discuss an admittedly selected set of issues that might be considered if and when the *Guidelines* are revised. In particular, we will review research in four areas that are particularly relevant to the *Guidelines*. After each section, we will describe the practical implications of this body of research. The four areas that we will discuss are the following (relevant sections of the *Guidelines* are included in parentheses):

Authors' Note: Correspondence regarding this article should be addressed to Wayne F. Cascio, Graduate School of Business Administration, University of Colorado at Denver, Campus Box 165, P.O. Box 173364, Denver, CO 80217-3364.

Review of Public Personnel Administration, Vol. 21, No. 3 Fall 2001 200-218

1. *Adverse impact, the four-fifths rule, and statistical significance testing*. We will discuss adverse impact and the four-fifths rule, including issues of statistical power and statistical significance testing (relevant to § 1607.4, "Information on Impact" and § 1607.14, "Technical Standards for Validity Studies").
2. *Criterion measures*. We will discuss properties of criterion measures, including dynamism of criteria, differences between typical and maximum performance, and multidimensionality of criteria (relevant to § 1607.4, "Technical Standards for Validity Studies", Point B[3], "Criterion Measures").
3. *Cutoff scores*. We will discuss professional issues, methods, and guidelines for setting cutoff scores, including the latest legal pronouncements in this area (relevant to § 1607.5, "General Standards for Validity Studies", Point H, "Cutoff Scores").
4. *Differential prediction*. We will discuss recent findings in the area of differential prediction (i.e., fairness analysis) and factors affecting the accuracy of conclusions based on differential prediction analysis (relevant to § 1607.14, "Technical Standards for Validity Studies", Point B[8], "Fairness").

Another way to view the above outline is in terms of practical questions related to the implementation of the *Guidelines* (1978). That is, one begins by asking, Is there a basis for enforcement of the *Guidelines* (our Section 1)? If yes, is there overall evidence of validity? Specifically, if an empirical research design is used, is there justification for the measures of performance, that is, criterion measures, used (our Section 2)? Is there justification for any cutoff scores used (our Section 3)? and If there is overall evidence of validity, is there specific evidence of unfairness for subgroups (our Section 4)?

ADVERSE IMPACT, THE FOUR-FIFTHS RULE, AND STATISTICAL SIGNIFICANCE TESTING

Section 1607.4, "Information on Impact", does not include any discussion of statistical power or statistical significance testing in evaluating differences in pass rates or selection rates from two or more subgroups of individuals. On this issue, the *Guidelines* (1978) recommends an arbitrary rule of thumb known as the four-fifths rule. This rule states that a difference in pass rates between two subgroups is not generally considered substantial if the pass rate for one subgroup is at least four-fifths (80%) of the pass rate for the higher subgroup. As Shoben (1978) noted,

> The four-fifths rule is an ill-conceived resolution of the problem of assessing the substantiality of pass or acceptance rate differences. It will produce anomalous results in certain cases because it fails to take account of differ-

> ences in sampling size. It also neglects the magnitude of differences in pass rates by considering only the ratio of the two rates. (pp. 805-806)

Shoben (1978) argued that the flaws in the four-fifths rule can be eliminated by replacing it with a test of the statistical significance of differences in pass rate proportions. Such a test takes into consideration the size of the sample and the magnitude of the differences in pass rates.

In fact, a disparity in pass or acceptance rates has one of three possible causes, as the United States Court of Appeals (D.C. Circuit) in *Palmer v. Shultz* (1987) pointed out. One, the disparity may be a product of unlawful discrimination. Two, the disparity may have a legitimate and nondiscriminatory cause. Three, the disparity may simply be a product of chance.

> A statistical analysis of a disparity in selection rates can reveal the probability that the disparity is merely a random deviation from perfectly equal selection rates. Statistics, however, can not entirely rule out the possibility that chance caused the disparity. Nor can statistics determine, if chance is an unlikely explanation, whether the more probable cause was intentional discrimination or a legitimate non-discriminatory factor in the selection process (p. 11, italics in original).
>
> Title VII nevertheless provides that if the disparity between selection rates . . . is sufficiently large so that the probability that the disparities resulted from chance is sufficiently small, then a court will infer from the numbers alone that, more likely than not, the disparity was a product of unlawful discrimination—unless the defendant can introduce evidence of a nondiscriminatory explanation for the disparity or can rebut the inference of discrimination in some other way. (pp. 11-12)

The court (*Palmer*, 1987) recommended the use of a .05 level of statistical significance, and two-tailed tests, in which "a statistically significant deviation in either direction from an equality in selection rates would constitute a prima facie case of unlawful discrimination" (p. 21).

In spite of the court's endorsement, null hypothesis significance testing (e.g., a test of a null hypothesis of equality of selection rates across subgroups) has been, and still is, a topic of heated debate in the scientific community (e.g., Aguinis, 1995; Chow, 1988, 1996; Cohen, 1994; Cortina & Folger, 1998; Murphy, in press; Murphy & Myors, 1998). Researchers have written extensively on the purpose, meaning, and use of significance testing. Some argue that significance testing is useful (e.g., Wainer, 1999), whereas others believe that it is misleading and should be discontinued (e.g., Schmidt, 1996). Next, we frame the issues of meaning, purpose, and use of significance testing within the context of adverse impact analysis (i.e.,

a significance test of the null hypothesis that the selection rates are equal across subgroups).

Purpose of Significance Testing

The purpose of significance testing is to determine whether a finding of inequality of selection rates in a sample of applicants can be explained by chance alone (i.e., sample fluctuations). Significance testing is used only when we use samples of applicants to make inferences regarding populations of applicants. Significance testing is not needed when we do not wish to make inferences from samples to populations. For instance, assume the admittedly unrealistic situation in which we use a selection instrument one time only with one sample of applicants only. Assume that Subgroup 1 (e.g., men) has a selection rate of .50 (i.e., 50% of men are given a job offer), and Subgroup 2 (e.g., women) has a selection rate of .60 (i.e., 40% of women are given a job offer). The conclusion is that these rates are different. In other words, men are selected at a greater rate as compared to women (i.e., 50% > 40%). Now assume the more typical situation in which we have a sample of applicants, but we are planning on using the selection procedure in the future with other applicants. In this case, we need to infer whether the .10 difference in the selection rates in our sample can be explained by chance alone (i.e., sample fluctuations) or by a robust finding. In this situation, in which we make inferences from samples to populations, the purpose of significance testing is to provide information regarding whether the .10 difference in sample selection rates can be explained by likely differences in the populations or by chance alone.

Meaning of Significance Testing

Rejecting the null hypothesis means that the hypothesis of equality of selection rates across subgroups is likely to be false. However, rejecting this null hypothesis does not inform us about the magnitude of this difference. Referring back to the above example of a difference between selection rates for women and men, rejecting this null hypothesis says nothing regarding the size of the difference in selection rates. All we can infer is that based on sample information, there is a difference in selection rates in the populations. However, significance testing does not allow us to make a statement regarding how large this difference is and what causes it.

Not rejecting the null hypothesis means that even if there are observed differences in selection rates across subgroups in the sample, we cannot rule out the possibility that the sample difference is due to chance alone and, in fact, selection rates are equal in the populations. In addition, not rejecting the null hypothesis does not mean that the selection rates are equal in the population. It just means that based on the sample information we have, there is not sufficient empirical evidence to conclude that differences exist. In fact, it is possible that there may be differences in the population, but these differences may be undetected in the sample, mainly due to inadequate statistical power (Morris & Lobsenz, 2000). Statistical power is defined as the probability that given the existence of population differences, these differences will be detected in the sample. Unless large samples are used, statistical power is likely to be insufficient to detect differences, and population inequalities in selection rates may go undetected.

Use of Significance Testing

The controversies surrounding significance testing seem to be due mainly to how significance testing is used. Stated differently, many researchers have noted that significance testing is abused and misused (e.g., Cohen, 1994; Schmidt, 1996). Significance testing allows us to infer whether the null hypothesis that selection rates are equal in the population is likely to be false. On the other hand, significance testing is incorrectly used when: (a) conclusions are made regarding the magnitude of selection rate differences across subgroups (e.g., a statistically significant result at the .01 level is interpreted as a larger difference than a result at the .05 level), and (b) failure to reject the null hypothesis is interpreted as evidence of lack of differences in selection rates in the population (i.e., not detecting differences in the sample may be due to insufficient statistical power).

Implications for Practice

The *Guidelines* (1978) do not include an in-depth discussion of uses and misuses of significance testing and the role of statistical power in assessing potential adverse impact (i.e., inequality of selection rates across subgroups). Moreover, the *Guidelines* recommend the arbitrary four-fifths rule of thumb without considering that insufficient statistical power may lead to the incorrect sample-based inference that there is no adverse impact in the

populations. Recent advances suggest that the assessment of adverse impact may benefit from significance testing. However, human resources (HR) officers should be aware of the purpose, meaning, and use of significance testing in determining whether selection instruments produce unequal selection rates across subgroups.

CRITERION MEASURES

The *Guidelines* (1978) provide several recommendations regarding the use of criterion measures (i.e., measures of job performance or other outcomes used to validate selection procedures empirically). For example, Section 1607.14-B ("Technical Standards for Validity Studies," Point 3), recommends that criteria represent important or critical work behavior(s) or work outcomes including, but not limited to, a standardized rating of overall work performance, performance in training (e.g., instructor evaluations, performance samples, and tests), production rate, error rate, tardiness, and absenteeism. Criterion measures are closely reviewed for job relevance, particularly measures consisting of paper-and-pencil tests.

If the criterion measure(s) used in conducting a validity study are deficient (i.e., important work behaviors and outcomes are not included in the measure) or contaminated (i.e., irrelevant work behaviors and outcomes are included in the measure), the results of a validity study do not provide useful information regarding the selection procedure. In addition to the well-known deficiency and contamination issues, researchers in the field of I&O psychology have investigated the following phenomena that have important implications for the selection of criterion measures and the conduct of validation studies: (a) dynamism of criteria, (b) distinction between typical and maximum performance, and (c) multidimensionality of criteria. We discuss each of these issues next.

Dynamism of Criteria

Almost half a century ago, Ghiselli (1956) discussed various features of criteria and noted that some criteria may be dynamic. More recently, Barrett, Caldwell, and Alexander (1985) suggested that dynamic criteria might assume one of the following three possible forms: (a) changes over time in average levels of group performance, (b) changes in validity coefficients over time, and (c) changes in the rank ordering of scores on the crite-

rion over time. The third form of dynamic criteria (i.e., changes in rank order of individuals over time) has attracted the attention of I&O psychologists (e.g., Hofmann, Jacobs, & Baratta, 1993; Hulin, Henry, & Noon, 1990) because of the implications for the conduct of validation studies and personnel selection in general. If the rank ordering of individuals on a criterion changes over time, future performance becomes a moving target. Under those circumstances, it becomes progressively more difficult to predict performance accurately the farther out in time from the original assessment. Are criteria really dynamic? In other words, do performance levels show systematic fluctuations across individuals? The answer seems to be in the affirmative.

Deadrick and Madigan (1990) collected weekly performance data from three samples of sewing machine operators (i.e., a routine job and a stable work environment). Results showed that the correlations between performance measures over time were smaller when the time lags increased (e.g., the correlation between Month 1 and Month 2 was greater than the correlation between Month 1 and Month 5). Deadrick and Madigan concluded that relative performance is not stable over time. A similar conclusion was reached by Hulin et al. (1990) and Hofmann et al. (1993): Individuals do tend to change their rank order of performance over time. A second issue regarding criterion measures is typical versus maximum performance.

Distinction Between Typical and Maximum Performance

Sackett, Zedeck, and Fogli (1988) distinguished typical performance from maximum performance (see also DuBois, Sackett, Zedeck, & Fogli, 1993). Typical performance refers to the average level of an employee's performance, whereas maximum performance refers to the peak level of performance an employee can achieve. In a study involving employees working in a large organization, Sackett et al. found that employees were more likely to perform at maximum levels when their performance was, to their knowledge, closely scrutinized. On the other hand, they performed at typical levels when they were not aware that their performance was being monitored. Moreover, results of this study demonstrated that measures of maximum performance (i.e., what employees can do) correlate only slightly with measures of typical performance (i.e., what employees will do). A final consideration regarding criterion measures is the multidimensionality of criteria.

Multidimensionality of Criteria

Researchers in I&O psychology have long recognized that job performance is a multidimensional construct (e.g., Schmidt, & Kaplan, 1971). Consequently, measures of performance (i.e., criterion measures) ought also to be multidimensional.

Campbell, McCloy, Oppler, & Sager (1993) described eight dimensions of performance believed to be comprehensive enough to describe all the jobs included in the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1991). Borman and Motowidlo (1997) proposed a simpler, two-dimensional taxonomy: task performance and contextual performance. Task performance is defined as (a) activities that transform raw materials into the goods and services that are produced by the organization and (b) activities that help with the transformation process by replenishing the supply of raw materials, distributing its finished products, or providing important planning, coordination, supervising, or staff functions that enable it to function effectively and efficiently. Contextual performance is defined as those behaviors that contribute to the organization's effectiveness by providing a good environment in which task performance can occur. Contextual performance includes behaviors such as

- persisting with enthusiasm and exerting extra effort as necessary to complete one's own task activities successfully (e.g., being punctual and rarely absent, expending extra effort on the job);
- volunteering to carry out task activities that are not formally part of the job (e.g., suggesting organizational improvements, making constructive suggestions);
- helping and cooperating with others (e.g., assisting and helping coworkers and customers);
- following organizational rules and procedures (e.g., following orders and regulations and respect for authority, complying with organizational values and policies); and
- endorsing, supporting, and defending organizational objectives (e.g., organizational loyalty, representing the organization favorably to outsiders).

Implications for Practice

The dynamism of criteria has several meaningful implications for practice. First, researchers should attempt to identify and understand the variables that cause differences in patterns of performance change over time.

For instance, some individuals may learn a job faster than others, and individuals may differ in self-efficacy, need for achievement, or self-esteem. A better understanding of the impact of each of these variables for specific jobs will allow for the development of better criterion measures. Second, given the multidimensionality of criteria it may be that for specific jobs, some dimensions are more dynamic than others. A better understanding of which performance dimensions are more likely to be dynamic will also allow for the development of better criterion measures. That is, prediction is likely to be more accurate for the less dynamic (more stable) dimensions.

Research regarding the distinction between typical and maximum performance has implications for the use of criterion measures in validation studies. Selection procedures are commonly administered in environments conducive to maximum performance (i.e., applicants are aware their performance is being monitored and the assessment of performance takes place over a short period of time). On the other hand, criterion measures are commonly administered in environments conducive to typical performance (e.g., employees are not always aware that their performance is being observed, and supervisors observe job-related behaviors over a long period of time). Thus, there is a lack of congruence between the performance construct assessed by selection procedures (i.e., maximum performance) and the performance construct assessed by criterion measures (i.e., typical performance). This lack of congruence may explain, at least in part, the difficulty in developing selection procedures accounting for more than 25% of the variance in performance scores (i.e., $r \cong .50$). Furthermore, the choice of a specific criterion measure in a validation study needs to consider whether scores are likely to be predicted by a selection procedure targeting typical or maximum performance.

Finally, based on the Cambpell et al. (1993) and Borman and Motowidlo (1997) taxonomies of performance, it becomes evident that choosing appropriate criteria in conducting a validation study can be more complex than implied in the *Guidelines* (1978). A practical implication of this discussion is that, at a minimum, criterion measures should include both task-specific and non-task-specific dimensions. In addition, a multiple-dimension situation suggests that more advanced data-analytic approaches may be needed (e.g., Murphy & Shiarella, 1997). In fact, in today's public administration environment, where technology requires constant learning of new tools in a cooperative environment, changes in organizational structure require the ability to work in teams, greater cross-agency mobility requires knowledge beyond one's specific tasks, and pressure to improve customer service requires interpersonal skills and abilities,

it could be argued that non-task-specific performance may be at least as important as task-related performance. Thus, good measures of non-task-specific performance ought to be developed so they can be used as yardsticks to evaluate the validity of selection procedures.

CUTOFF SCORES

Although there are instances where cutoff scores need not be set, as when rank-order (top-down) hiring is used, civil service rules frequently require that a cut score be established to determine who passed and who did not pass an entry-level or promotional examination. How does one actually set a cut score? Various sets of professional standards and guidelines provide broad, general guidance on this issue. The following are excerpts from three such guidelines.

Section 1607.5-H, "General Standards for Validity Studies" (cutoff scores), from the *Guidelines* (1978), notes that

> where cutoff scores are used they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the workforce. Where applicants are ranked on the basis of properly validated selection procedures and those applicants scoring below a higher cutoff score than appropriate in light of such expectations have little or no chance of being selected for employment, the higher cutoff score may be appropriate, but the degree of adverse impact should be considered.

Likewise, the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 1987) specify that

> cutoff or other critical scores may be set as high or as low as the purposes of the organization require, if they are based on valid predictors. This implies that (a) the purposes of selection are clear and (b) they are acceptable in the social and legal context in which the employing organization functions. (p. 32)

Finally, the *Standards for Educational and Psychological Testing* (American Education Research, 1999) address issues of cutoff scores in several sections. For example,

> cut scores may be established to select a specified number of examinees (e.g., to fill existing vacancies) in which case little further documentation may be needed concerning the specific question of how the cut scores are estab-

> lished, though attention should be paid to the legal requirements that may apply. (Standard 4.19)

Standard 4.21 notes that

> cut scores are sometimes based on judgments about the adequacy of item or test performances . . . or performance levels (e.g., the level that would characterize a borderline examinee). The procedures used to elicit such judgments should result in reasonable, defensible standards that accurately reflect the judges' values and intentions.

Professional guidelines and standards often caution users to be aware of legal requirements that may apply. An important appellate court decision in *Lanning v. Southeastern Pennsylvania Transportation Authority* (1999), established the legal standard to apply when evaluating an employer's business justification in response to a challenge to that employer's cutoff score on a pre-employment assessment procedure. Previously, the Civil Rights Act of 1991 defined the employer's rebuttal to a "disparate impact" discrimination claim involving objective assessment as "job related for the position in question and consistent with business necessity" (§ 105).

The court emphasized that tests for job relatedness and business necessity comprise two separate tests. Job relatedness may be shown, for example, by demonstrating a statistically significant relationship between test scores and a measure of job performance (i.e., criterion-related validity). However, when a cutoff score produces a disproportionate impact against a protected subgroup and it is challenged, the court wrote, "The business necessity prong must be read to demand an inquiry into whether the score reflects the minimum qualifications necessary to perform successfully the job in question." Hence, "business necessity" requires setting a cutoff score that reflects the minimum standard necessary to perform a job successfully.

How do the guidelines and standards referenced earlier fit with this legal standard? The *Guidelines* (1978) emphasize that cutoff scores should be set so as to be consistent with normal expectations of acceptable proficiency within the work force. If one reads that statement to imply minimally acceptable proficiency, then the *Guidelines* would be consistent with the *Lanning* decision that emphasizes the setting of a cutoff that reflects minimum qualifications. The SIOP *Principles* (1987) emphasize the job-relatedness test, that is, the demonstration of evidence of validity, but then leaves the actual setting of a cut score to the discretion of decision makers. This ignores the business necessity test that the *Lanning* decision requires. Finally, the *Standards* (American Educational Research, 1999) acknowledges the fact that cut

scores may be set simply to correspond to the number of vacancies an employer has to fill, or they may reflect minimum qualifications. The latter approach is consistent with the *Lanning* decision.

Setting Minimum Standards

One method for setting minimum standards follows the Angoff (1971) procedure. In this approach, expert judges rate each item in terms of the probability that a barely or minimally competent person would answer the item correctly. The probabilities (or proportions) are then averaged for each item across judges to yield item cutoff scores, and item cutoff scores are summed to yield a test cutoff score. The method is easy to administer, it is as reliable as other judgmental methods for setting cutoff scores, and it has intuitive appeal because expert judges (rather than a consultant) use their knowledge and experience to help determine minimum performance standards. Not surprisingly, therefore, the Angoff method has become the favored judgmental method for setting cutoff scores on employment tests (Cascio, Alexander, & Barrett, 1988; Maurer & Alexander, 1992). If the method is to produce optimal results, however, judges should be chosen carefully based on their knowledge of the job and the knowledge, skills, abilities, and other characteristics needed to perform it. Then, they should be trained to develop a common conceptual framework of a minimally competent person (Maurer & Alexander, 1992; Maurer, Alexander, Callahan, Bailey, & Dambrot, 1991). Finally, it is important to recognize that if a test consists of items that most of the judges can answer correctly, then judges may make higher Angoff judgments when provided with answers to test items. The result may be a test with a higher cutoff score than that obtained when judges are not provided with answers (Hudson & Campion, 1994).

Implications for Practice

The *Standards* (American Educational Research, 1999) recommends the following:

> If a judgmental standard-setting process is followed, the method employed should be clearly described, and the precise nature of the judgments called for should be presented, whether those are judgments of persons, of item or test performances, or of other criterion performances predicted by test scores. Documentation should also include the selection and qualification

> of judges, training provided, any feedback to judges concerning the implications of their provisional judgments, and any opportunities for judges to confer with one another. Where applicable, variability over judges should be reported. Whenever feasible, an estimate should be provided of the amount of variation in cut scores that might be expected if the standard-setting procedure were replicated. (Standard 4.19)

These recommendations are sound no matter which specific method of setting cutoff scores decision makers use.

In addition to these recommendations, we make the following 8 recommendations for practice on the basis of two reviews of the literature on cutoff scores (Cascio et al., 1988; Truxillo, Donahue, & Sulzer, 1996), and in light of the *Lanning* decision.

1. Determine if it is necessary to set a cutoff score at all; legal and professional guidelines do not demand their use in all situations.
2. It is unrealistic to expect that there is a single best method of setting cutoff scores for all situations.
3. Begin with a job analysis that identifies relative levels of proficiency on critical knowledge, skills, abilities, and other characteristics.
4. The validity and job relatedness of the assessment procedure are critical considerations.
5. If a cutoff score is to be used as an indicator of minimum proficiency, relating it to what is necessary on the job is essential. Normative methods of establishing a cut score (in which a cut score is set based on the relative performance of examinees) do not indicate what is necessary on the job.
6. When using judgmental methods, sample a sufficient number of judges, for example, 7 to 10.
7. Consider statistical (standard error of measurement), and legal (adverse impact) issues when setting a cut score.
8. Set cutoff scores high enough to ensure that minimum standards of job performance are met.

DIFFERENTIAL PREDICTION

The *Guidelines* (1978) defines fairness as a situation when "members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than do members of another group, and the differences in scores are not reflected in differences in a measure of job performance" (§ 1607.14, Point B[8]). Stated differently, unfairness occurs when the relationship between selection procedure scores and job performance scores differs across subgroups. Because the relationship between test scores and

performance changes, or is moderated by, the group-membership variable, the variables race, sex, and ethnic group are labeled *moderators*.

Typically, fairness, also called *differential prediction*, is assessed by using moderated multiple regression (MMR) (Bartlett, Bobko, Mossier, & Hannan, 1978; Cleary, 1968). This involves a regression equation including Y (e.g., supervisory ratings) as a criterion, X (selection procedure scores) as a predictor, and Z (e.g., sex-coded 1 for women and 2 for men) as a second predictor. In addition, the MMR equation includes a third predictor consisting of the $X \bullet Z$ product. This product term carries information regarding the potential moderating effect of Z on the X-Y relationship. The MMR equation is the following:

$$\hat{Y} = a + b_1X + b_2Z + b_3X \bullet Z, \quad (1)$$

where $\hat{Y}$ is the predicted value for Y, a is the least-squares estimate of the intercept, b_1 is the least-squares estimate of the population regression coefficient for X, b_2 is the least-squares estimate of the population regression coefficient for Z, and b_3 is the least-squares estimate of the population regression coefficient for the product term that carries information about the moderating effect of Z (Cohen & Cohen, 1983). Rejecting the null hypothesis that β_3 (i.e., the population value of b_3) = 0 indicates the presence of a moderating effect (i.e., unfairness). Stated differently, rejecting this null hypothesis indicates that the selection procedure does not predict performance equally well for the subgroups under consideration.

Research conducted over the past few years has revealed that conclusions regarding fairness of selection procedures based on MMR may suffer from a serious problem. More precisely, MMR analyses are typically conducted at low levels of statistical power (for a review, see Aguinis, 1995). In practical terms, low power affects personnel testing in that one may incorrectly conclude that a selection procedure predicts performance equally well for various subgroups based on race or sex. However, this sample-based conclusion may be incorrect. In fact, the selection procedure may predict performance differentially across subgroups and, consequently, be unfair to members of protected classes. Next, we review factors that adversely affect the statistical power of MMR and, therefore, may cause HR officers to use selection procedures that are unfair to members of specific subgroups.

Variance Reduction in Test Scores

The power of MMR is reduced markedly when the variance of test scores (i.e., *X*) is smaller in the sample than in the population (Aguinis & Stone-Romero, 1997). This reduction in variance, also labeled *direct range restriction*, is typical in criterion-related validity studies. Decisions regarding which individuals to select for an opening are frequently based on their standing on a predictor variable *X* (e.g., test of job aptitude); only those who obtain a score that exceeds a specific cutoff point are selected, leading to an *X* variance in the sample that is smaller than the *X* variance in the population. Moreover, although Aguinis and Stone-Romero investigated direct range restriction, other forms of range restriction are also pervasive in personnel selection (Aguinis & Whitehead, 1997) and have a detrimental effect on statistical power.

Sample Size Heterogeneity

Typically, in personnel selection, there are unequal sample sizes across the levels of *Z* (e.g., more Whites than Latinos and African Americans). As a consequence of this situation, the statistical power to detect ethnicity or gender as a moderator variable is reduced, and conclusions regarding fairness analysis may be erroneous.

An empirical examination of this issue using Monte Carlo simulations (Stone-Romero, Alliger, & Aguinis, 1994) demonstrated the effect of unequal sample sizes across moderator-based subgroups on the power of MMR. In situations with two subgroups (e.g., ethnicity coded as majority vs. minority), results showed that there was a considerable decrease in power when the size of Subgroup 1 was .10 relative to total sample size regardless of total sample size. A proportion of .30, closer to the optimum value of .50, also reduced the statistical power of MMR but to a lesser extent.

Error Variance Heterogeneity

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 subgroup, the error variance is estimated by the mean square residual from the regression of *Y* on *X*:

$$\sigma^2_{e(i)} = \sigma_{Y(i)}(1-\rho^2_{XY(i)}), \quad (2)$$

where $\sigma_{Y(i)}$ and $\rho_{XY(i)}$ are the *Y* standard deviation and the *X*-*Y* correlation in each subgroup, 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.

Heterogenous error variances can affect both Type I (incorrectly concluding that the selection procedures are unfair) error 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. As noted in Aguinis and Pierce's (1998a) review, 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.

Implications for Practice

Based on the above discussion of recent findings, we suggest that HR officers use computer programs to calculate the power of their fairness analyses before concluding that personnel selection procedures are fair. Such programs are in the public domain. Descriptions of them, as well as instructions on how to obtain them, can be found in Aguinis, Pierce, and Stone-Romero (1994), Aguinis and Pierce (1998b), and Aguinis, Boik, and Pierce (in press). In addition, we recommend that researchers use a Web-based program described by Aguinis, Petersen, and Pierce (1999) to compute alternatives to MMR when the homogeneity-of-error assumption is violated.

CONCLUDING REMARKS

The purpose of the present paper was to provide an update and implications for practice regarding an admittedly selected set of advances in the field of I&O psychology relevant to various sections of the 1978 *Guidelines*. Our discussion reveals that recent research-based advances regarding (a) adverse impact, the fourth-fifths rule, and significance testing; (b) crite-

rion measures; (c) cutoff scores; and (d) differential prediction have meaningful implications for the development and use of selection procedures vis-à-vis the *Guidelines'* recommendations. First, procedures implemented to investigate adverse impact ought to consider the issue of statistical power as well as the meaning, purpose, and use of significance testing. Second, the development of criterion measures ought to consider the dynamism of criteria, the distinction between typical and maximum performance, and the multidimensionality of criteria. Third, HR officers ought to consider that setting cutoff scores involves several steps beginning with a determination of whether it is necessary to establish a cutoff score at all. However, if a cutoff score is set, and it leads to an adverse impact on one or more protected groups, it should reflect minimum qualification standards. Fourth, assessing the fairness/unfairness of selection procedures ought to consider recent findings regarding the low statistical power of the differential prediction test and, consequently, the possibility that a conclusion that a selection procedure is fair may not be warranted.

In closing, recent research-based advances described in the present article may serve as useful input if and when the 1978 *Guidelines* are revised. In the meantime, we hope they will provide HR officers with useful information as they strive to develop and use selection procedures that are valid and fair.

REFERENCES

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

Aguinis, H., Boik, R. J., & Pierce, C. A. (in press). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods.*

Aguinis, H., Petersen, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods*, *2*, 315-339.

Aguinis, H., & Pierce, C. A. (1998a). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods*, *1*, 296-314.

Aguinis, H., & Pierce, C. A. (1998b). Statistical power computations for detecting dichotomous moderator variables with moderated multiple regression. *Educational and Psychological Measurement*, *58*, 668-676.

Aguinis, H., Pierce, C. A., & Stone-Romero, E. F. (1994). Estimating the power to detect dichotomous moderators with moderated multiple regression. *Educational and Psychological Measurement*, *54*, 690-692.

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.

Aguinis, H., & Whitehead, R. (1997). Sampling variance in the correlation coefficient under indirect range restriction: Implications for validity generalization. *Journal of Applied Psychology*, *82*, 528-538.

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, *115*, 308-314.

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.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In Thorndike, R. L. (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Barrett, G. G., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology*, *38*, 41-56.

Bartlett, C. J., Bobko, P., Mossier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated regression strategy: An alternative to differential analysis. *Personnel Psychology*, *31*, 223-241.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance*, *10*, 99-109.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel psychology in organizations* (pp. 35-70). San Francisco: Jossey-Bass.

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology*, *41*, 1-24.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*, 105-110.

Chow, S. L. (1996). *Statistical significance: Rationale, validity, and utility*. Thousand Oaks, CA: Sage.

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

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, *49*, 997-1003.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods*, *1*, 334-350.

Deadrick, D. L., & Madigan, R. M. (1990). Dynamic criteria revisited: A longitudinal study of performance stability and predictive validity. *Personnel Psychology*, *43*, 717-744.

DuBois, C. L., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and White-Black differences. *Journal of Applied Psychology*, *78*, 205-211.

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology*, *40*, 1-4.

Hofmann, D. A., Jacobs, R., & Baratta, J. E. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology*, *78*, 194-204.

Hudson, J. P., Jr., & Campion, J. E. (1994). Hindsight bias in an application of the Angoff method for setting cutoff scores. *Journal of Applied Psychology*, *79*, 860-865.

Hulin, C. L., Henry, R. A., & Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin*, *107*, 328-340.

Lanning v. Southeastern Pennsylvania Transportation Authority, No. 98-1644 (3d Cir. filed June 29, 1999).

Maurer, T. J., & Alexander, R. A. (1992). Methods of improving employment test critical scores derived by judging test content: A review and critique. *Personnel Psychology*, *45*, 727-762.

Maurer, T. J., Alexander, R. A., Callahan, C. M., Bailey, J. J., & Dambrot, F. H. (1991). Methodological and psychometric issues in setting cutoff scores using the Angoff method. *Personnel Psychology*, *44*, 235-262.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology*, *53*, 89-111.

Murphy, K. R. (in press). Using power analysis to evaluate and improve research. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology.* Malden, MA: Blackwell Publishers.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum.

Murphy, K. R., & Shiarella, A. H. (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.

Palmer v. Shultz, No. 85-6101 (U.S. App. D.C. filed March 24, 1987).

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*, 482-486.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology*, *24*, 419-434.

Shoben, E. W. (1978). Differential pass rates in employment testing: Statistical proof under Title VII. *Harvard Law Review*, *91*, 793-813.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Stone-Romero, E. F., Alliger, G. M., & Aguinis, H. (1994). Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management*, *20*, 167-178.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting cutoff scores for personnel selection tests: Issues, illustrations, and recommendations. *Human Performance*, *9*, 275-295.

Uniform guidelines on employee selection procedures, 43 Fed. Reg. 38290-38315 (1978).

U.S. Department of Labor. (1991). *Dictionary of occupational titles* (4th ed.). Washington, DC: Author.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212-213.