Journal of Applied Psychology
1985, Vol. 70, No. 1, 56-65

0021-9010/85/$00.75

A Meta-Analysis of Ratee Race Effects in Performance Ratings

Kurt Kraiger
University of Colorado at Denver

J. Kevin Ford
Michigan State University

A recent review of ratee race effects on performance ratings (Landy & Farr, 1980) found conflicting results. For the present research, meta-analytic techniques were used for more substantive conclusions about the existence of ratee race effects and whether the effects were related to rater race or were moderated by situational factors. The five moderators examined included the study setting (laboratory/field), rater training (offered/not offered), type of rating (behavior/trait), rating purpose (administrative/research), and the composition of the workgroup (percentage of blacks in each study). Seventy-four studies with a total sample of 17,159 ratees were located for white raters, whereas 14 studies with 2,428 ratees included data on black raters. The corrected mean correlations between ratee race and ratings for white and black raters were .183 and −.220, respectively, with 95% confidence intervals that excluded zero for both rater groups. Substantial moderating effects were found for study setting and for the saliency of blacks in the sample. Race effects were more likely in field settings when blacks composed a small percentage of the workforce. The appropriate role of laboratory studies and the implications of the results for guiding future research on racial bias are discussed.

A traditional literature review of ratee race effects has been recently completed by Landy and Farr (1980). Many of the studies reviewed showed a significant ratee race effect (deJung & Kaplan, 1962; Farr, O'Leary, & Bartlett, 1971; Hamner, Kim, Baird, & Bigoness, 1974; Landy & Farr, 1973), whereas other studies indicated no effect (Fox & Lefkowitz, 1974; Schmidt & Johnson, 1973). Some studies found effects for rater race (e.g., Campbell, Crooks, Mahoney, & Rock, 1973), whereas others did not (Schmidt & Johnson, 1973). Results of still other studies showed a more complex interaction between rater and ratee race and other variables such as ratee performance level (Bigoness, 1976), age (Toole, Gavin, Murdy, & Sells, 1972), and employment status (Bass & Turner, 1973). Landy and Farr (1980) concluded that, in general, ratees tend to receive higher ratings from raters of the same race, although situational factors may moderate this effect.

The nature of these diverse findings over various settings, populations, and operationalizations of measures makes any conclusions drawn from a traditional review tenuous at best. The conflicting results call for a more substantive strategy for analyzing the available data on ratee race effects. One such tool is meta-analysis, a method of integrating results from existing studies to reveal patterns of underlying relations and causalities (Hunter, Schmidt, & Jackson, 1982). The results from a meta-analysis not only provide a more definitive appraisal of race effects but allow the researcher to examine the viability of assumptions regarding conditions that may moderate the relationship between ratee race and rating. An examination of the existing rating literature suggests five potential moderators of ratee race effects.

The most recent viewpoint on race effects is that they are much more likely to be found in laboratory than "real world" field settings. Wendelken and Inn (1981) argued that in

This article is based on a paper presented at the 91st Annual Convention of the American Psychological Association, August, 1983, Anaheim, California.

The authors express their thanks to Neal Schmitt, Daniel Ilgen, and two anonymous reviewers for providing comments to earlier drafts of this article.

Requests for reprints should be sent to Kurt Kraiger, Department of Psychology, University of Colorado at Denver, 1100 Fourteenth Street, Denver, Colorado 80202.

typical laboratory studies, ratee race is made salient by artificially constraining ratee performance and by limiting the amount of information available to the rater. In contrast, the real world rater has a larger amount of job-relevant information available that reduces the necessity of using race as a factor in the evaluation. The contention that race is less salient in organizational settings is arguable. Under affirmative action and equal employment pressures, race may become a highly salient and rational consideration in the evaluation of performance in organizations (Mobley, 1982), motivating the rater, either intentionally or unintentionally, to minimize race effects. On the other hand, one could argue effectively that ratee race effects should be less likely in laboratory experiments due to the tendency of college students to provide socially desirable responses (Jones & Sigall, 1971). Consequently, it is unclear whether race effects should be more likely in laboratory or field settings.

A second issue is the impact of rater training on rating effects. Although a number of authors have suggested rater training has or should reduce race effects (e.g., Bigoness, 1976; Hamner, Kim, Baird, & Bigoness, 1974), previous training studies have focused almost exclusively on other systematic rating effects such as halo, leniency, and central tendency. Proponents of this view commonly cite one study by Schmidt and Johnson (1973) for support, but a careful reading of this study reveals that participants were not given any rater training but instead provided ratings after an extensive human relations course as part of a managerial training program. Furthermore, recent empirical evidence indicates that the impact of training on rating effects diminishes over time (Bernardin, 1978; Ivancevich, 1979) and often bears little relationship to the accuracy of ratings (Bernardin & Pence, 1980; Borman, 1979). Thus, although rater training potentially may reduce ratee race effects, there is no empirical evidence that it has, and reasonable cause for pessimism that it could.

A third issue is the extent to which ratee race effects are related to the composition of the workgroup. Although this proposition has not received much attention in the Industrial/Organizational (I/O) literature, it has been suggested that diminishing the proportion of minority to majority group members should increase the racial saliency of minority members and polarize evaluations (Taylor, 1981). For example, Schmitt and Hill (1977) reported that black female assessees tended to receive lower ratings when their assessment center group was composed principally of white males than when the group was more equally integrated by race and sex. The pervasiveness of the phenomenon across settings is not yet known.

A fourth issue concerns the behavioral specificity of the rating scale. Wherry and Bartlett (1982) have theorized that behaviorally based rating formats will be less prone to rating effects than more global, general formats. Thus, Brugnoli, Campion, and Basen (1979) reported that race effects were minimized when a behaviorally based rating scale was used instead of a global measure. Nevertheless, a comparison of rating formats by Jacobs, Kafry, and Zedeck (1980) found little evidence for the psychometric superiority of behavioral scales over alternative formats. Research on rating formats suggests varying rating formats will do little to reduce ratee race effects.

A final issue often discussed in the rating literature is the impact of the purpose or consequence of ratings on rating effects. As hypothesized by Wherry and Bartlett (1982), ratings obtained under experimental conditions should be more accurate than those obtained under actual job conditions where administrative actions may affect the rater or ratee. Relevant research has shown that ratings are significantly more lenient and exhibit more halo when raters are told the ratings are for administrative rather than research purposes (Sharon & Bartlett, 1969; Taylor & Wherry, 1951; Warmke & Billings, 1979). However, the impact of rating purpose on ratee race effects has not yet been explicitly studied.

In sum, the rating literature has suggested at least five potential moderators of race effects in performance evaluations. Existing empirical evidence for each potential moderator is often limited to sporadic single-setting correlates that may not be generalizable to other situations. The extent and operative nature of each potential moderator

can be investigated either by systematically varying the moderator variable across multiple organizational units or through the cumulation of the results of many studies employing different settings or conditions. Although the former option presents formidable cost and logistical obstacles, the later meta-analytic method is feasible and has already been successfully applied to other domains of I/O psychology (e.g., Schmidt, Hunter, & Pearlman, 1981; Terborg, Lee, Smith, Davis, & Turbin, 1982).

It should be noted at the outset that the present study does not directly isolate the important issue of racial bias in performance evaluations. As Wherry and Bartlett (1982) have stated, subjective ratings are a function of not only ratee performance but also biases in the observation and recall of that performance by the rater. Because individual studies rarely provide information regarding actual performance, a meta-analysis of race effects cannot separate the relative contributions of ratee performance and rater bias to rating differences. Nevertheless, a necessary first step to the investigation of racial bias is to determine whether race is related to evaluations under various conditions.

Thus, meta-analytic techniques are used for more substantive conclusions regarding the existence of ratee race effects and whether the race effects are related to rater race or are moderated by the five situational factors previously discussed.

Method

An attempt was made to locate, summarize, and analyze the results of all published studies and a number of unpublished studies reporting performance ratings of black and white ratees. The search was performed manually using three indexes (*Personnel Literature, Personnel Management Abstracts,* and *Psychological Abstracts*) and a systematic review of the *Journal of Applied Psychology* and *Personnel Psychology* from 1966 to 1981. Because misrepresentation of population parameters may result from inadequate sampling procedures (Hunter et al., 1982), a variety of unpublished works was included as well. Sources included solicited responses from researchers active in the areas of test validation and performance evaluation and technical reports readily available to the authors. A total of 34 published and 47 unpublished studies were located for analysis. These totals reflect multiple samples for some reports. Effect sizes were available or determinable from 30 published and 44 unpublished studies, resulting in a total sample size of 74. All studies presented data for white raters, but only 14 studies presented data for black raters. A complete list of the studies included in the analysis is presented in the Appendix.

Analysis

Different estimates of effect size have been proposed including d (Glass, 1976) and ω^2 (Hays, 1973) where ω^2 is calculated from t. However, both d and t are algebraic transformations of the more generally applied point-biserial correlation, r_{pb} (Hunter et al., 1982). The present meta-analysis cumulated the correlation of ratings and race (arbitrarily coded White = 1, Black = 0) in the computation of the mean effect size ($\bar{r}_{pb}$) and its variance (σ_{pb}^2) across studies. An estimate of variance due to sampling error (σ_e^2) and the population variance for effect sizes (σ_ρ^2) were computed using procedures explained in Hunter et al. (1982). The estimated standard error (σ_e) was used to establish confidence intervals around $\bar{r}_{pb}$ and to test the hypothesis that $\bar{r}_{pb} = 0$ in the population.

Because the size of a point-biserial correlation is affected by the relative proportions of the two groups, effect sizes for individual studies were corrected for differences in subgroup sample sizes (the corrected correlation estimates effect size if subgroup sample sizes were equal). Estimated sampling error was adjusted for added variance due to this correction. Cumulated effect sizes and estimated population variance were also corrected for the average reliability of ratings to estimate true effect size and variance given perfect measurement. The average reported conspect reliability across studies was .70. It should be noted that this value results in a more conservative correction than the .60 value advocated by Hunter & Schmidt (Hunter et al., 1982; King et al., 1980).

Estimated population variance (σ_ρ^2) represents actual study-to-study variation in effect sizes with estimated variance due to small samples and unequal sample sizes removed. This corrected variance may be trivial in size and likely due to other statistical artifacts or it may be nontrivial and suggest the possible presence of one or more moderators in the data. Two tests of the triviality in corrected variance were performed. The first was a chi-square test of the hypothesis of no variation suggested by Hunter et al., (1982). The second, less formal test was the Schmidt, Hunter, and Pearlman (1982) "bare bones" analysis, which computes the ratio of sampling error variance to true variance to determine whether the observed variance (over 75%) is largely artifactual in nature. The two tests for the triviality of the corrected variance revealed significant chi-squares, $\chi^2(74, N = 17{,}159) = 221.75$, $p < .01$, for white raters and $\chi^2(14, N = 2{,}428) = 76.41$, $p < .01$, for black raters, and a small sampling error to true variance ratio (less than 40%) for the total sample of studies. These results support the investigation for potential moderators in the data.

For white raters, five potential moderators were investigated. The small number of studies precluded the examination of moderator variables for black raters. Four moderators were coded as dichotomous variables: setting (lab/field), rater training (offered/not offered), type of rating (behavior based/trait), and rating purpose (administrative/research). The coding was completed primarily by the first author. A subset of studies was coded by both authors resulting in a high level (90%) of agreement. The

Table 1
Mean and Variance of Race Effects in Performance Ratings

	No. studies	Cumulated N	$\bar{r}_{pb}$	Corrected[a] $\bar{r}_{pb}$	Estimated d	σ_r^2	σ_ρ^2	Corrected confidence intervals	Z test[b]
Total sample									7.57**
White raters	74	17,159	.153	.183	.37	.012	.010	$.02 < \rho < .35$	
Black raters	14	2,428	−.184	−.220	−.45	.029	.032	$-.41 < \rho < -.03$	
White raters—moderator analyses									
Research setting									2.41*
Laboratory	10	1,010	.031	.037	.07	.021	.002	$-.20 < \rho < .27$	
Field	64	16,149	.160	.192	.39	.011	.008	$.03 < \rho < .35$	
Rater training[c]									1.32
Offered	16	5,055	.173	.207	.42	.013	.012		
Not offered	39	10,088	.158	.189	.38	.008	.005		
Rating scales[c]									1.12
Behavioral	32	8,692	.178	.213	.43	.011	.009		
Trait	23	6,451	.142	.170	.34	.007	.003		
Purpose[c]									1.19
Administrative	18	6,955	.166	.199	.41	.008	.006		
Research	37	8,259	.159	.191	.39	.010	.006		

[a] Correction for attenuation, average interrater reliability = .70.
[b] Z test is for subgroups immediately below test statistic.
[c] Variation by subgroup for field studies only.
* $p < .05$. ** $p < .001$.

presence of these moderators was examined by classifying the studies into relevant subsamples and recomputing subsample r_{pb}s and confidence intervals. Differences in subgroup effect sizes were tested for significance by a procedure adapted from Rosenthal and Rubin (1982). The procedure requires the definition of the subgroup comparison by a set of contrast weights such that the sum of the contrasts over all studies equals zero. The procedure also requires the computation of estimated variance of individual study effect sizes. Let L equal the sum of the product of contrast weights times effect sizes overall studies and M equal the sum of the contrasts over studies squared divided by the estimated effect size variance. Rosenthal and Rubin (1982) have shown that the quotient L/M is distributed as a standard normal deviate, Z.

The fifth moderator, racial composition of the workgroup, was investigated by coding the percentage of blacks rated in each study and correlating this value with the uncorrected effect size across all studies (i.e., the effect size before the correction for differences in subgroup sample size). A significant correlation across studies indicates the percentage of blacks rated moderates the degree of relationship between race and performance rating (Arnold, 1982). If significant, the correlation can be corrected for measurement error by the formula provided by Hunter et al. (1982).

Results

Effects for White and Black Raters

The major results of the meta-analysis are presented in Table 1. The table shows cumulated sample sizes, effect sizes, variance estimates, and confidence intervals for the performance ratings of black and white raters. Results of moderator analyses for white raters are also presented in Table 1.

The best estimate of the population effect size (ρ_{pb}) is the mean point-biserial correlation corrected for unreliability in ratings. For white raters, this estimate was .183 and was based on a sample of 74 studies and 17,159 ratees. The 95% confidence interval about this estimate excluded zero ($.02 < \rho < .35$), indicating that white raters assigned significantly higher ratings to white ratees than to black ratees. Transforming the corrected $\bar{r}_{pb}$ to Glass's (1976) d statistic (.37) suggests an alternate but compatible interpretation: The average white ratee received a more favorable evaluation from white raters than 64% of the black ratees. The estimated population variance of effect sizes for white raters ($\sigma_\rho^2 = .010$) was small but greater than zero and suggests the presence of one or more moderators in the data.

For black raters, the estimated population effect size (ρ_{pb}) was −.220 and was based on a sample of 14 studies and 2,428 ratees. The 95% confidence interval about this estimate

also excluded zero ($-.41 < \rho < -.03$), indicating that black raters assigned significantly higher ratings to black ratees than to white ratees. The d value for black raters ($-.45$) suggests that the average black ratee received a higher rating than 67% of the white ratees. Although the population variance of effect sizes for black raters ($\sigma_\rho^2 = .032$) was also greater than zero, moderator analyses were not performed because of the small number of studies available. A Z test by rater race (Rosenthal & Rosen, 1982) showed a highly significant difference in average effect sizes for black and white raters ($z = 7.57$; $p < .001$).

Moderator Analyses for White Raters

The effect of saliency of blacks on rating differences was tested by computing the correlation across all studies of the uncorrected study effect size and the percentage of black ratees. The correlation was found to be $-.16$ ($N = 73$; $p < .10$) or $-.18$ when corrected for measurement error. The negative sign indicates that effect size increases as the percentage of blacks decreases.

The results of the tests for the four dichotomous moderators are presented in Table 1. Effect sizes found in field setting ($\bar{r}_{pb} = .192$) were significantly larger than effect sizes found in laboratory settings ($\bar{r}_{pb} = .037$; $z = 2.41$; $p < .05$). Because the 95% confidence intervals for laboratory studies includes zero, the hypothesis of no race differences cannot be rejected.

Due to the large discrepancy between laboratory and field results, further tests for moderators were completed with both the entire sample and field studies alone. The results for both sets of studies revealed similar patterns. Theoretically, the results for field studies alone should be more relevant because the remaining moderators (rater training, rater format, and rating purpose) represent organizational conditions rather than psychological processes that may occur in any setting. Therefore, tests of potential moderators for field studies are presented. Examination of the field data revealed no significant differences in subgroup effect sizes for rater training, rating format, or rating purposes. Though nonsignificant, race effects were slightly higher for trained raters than for untrained raters and for ratings made with behavioral scales versus trait scales.

Discussion

An examination of the results indicates that effect sizes for white raters were positive in 68 of the 74 studies but significantly greater than zero in only 46 cases. A traditional literature review might label such results as conflicting or mixed. The meta-analysis results provide a more definitive conclusion regarding ratee race effects and potential moderator variables.

The size of the ratee race effect is somewhat small for both black and white raters but is relatively consistent across studies. Of greater importance is the difference in the direction of the effect for the two sets of raters. Both black and white raters gave significantly higher ratings to members of their own race. This evidence concurs with statements made by Landy and Farr (1980) and lead us to conclude that race of ratee does have an impact on performance ratings in real world settings. This tendency is equally strong for both black and white raters.

The investigation of potential moderators for white raters demonstrated that effect size was not influenced by type of rating, rating purpose, or rater training. The absence of a moderating effect for type and purpose of rating is contrary to prevalent theory in the performance appraisal domain (e.g., Wherry & Bartlett, 1982). Behaviorally based ratings appear equally prone to race effects as trait ratings. The results are consistent with Landy and Farr's (1980) conclusion that rating formats account for little variance in ratings. Interestingly, Feldman (1981) has noted that the cognitive processes of raters involving the observation, categorization, and explanation of behavior may be impervious to changes in rating format. In other words, raters may selectively attend to and recall behaviors that validate their underlying global (trait) impressions.

The lack of moderating effect for rating purpose was somewhat unexpected because purpose has been found to affect systematic rating effects such as halo and leniency. It appears that the impact of rating purpose does not generalize to ratee race effects. This

reinforces Warmke and Billings's (1979) call for more emphasis on understanding how organizational contextual variables such as rating purpose affect rating criteria.

The lack of an effect for rater training may reflect the fact that rater training programs rarely address issues of differential treatment of ratee subgroups. From this perspective, it could be argued that the data do not allow for an adequate test of the potential impact of rater training on race effects. Alternatively, the results can be viewed as consistent with recent reviews (cf. Bernardin & Pence, 1980) suggesting that rater training has minimal impacts on rating effects. At the very least, the results cast doubt on the rise of traditional rater training programs to reduce ratee race effects in ratings.

A substantial moderating effect was found for the study setting. Race effects, while substantial in field settings, were negligible in laboratory studies. This finding directly contradicts assertions by Wendelken and Inn (1981) and others that race effects are minimized in the field as opposed to laboratory settings. One explanation for this finding is that researchers can ensure equal performance among subgroups in laboratory but not field settings. An alternative (but not contradictory) explanation is that any tendency of subjects to evaluate blacks and whites differentially in laboratory settings is overwhelmed by their desire to give socially desirable responses.

As an illustration, a recent study of sex effects in the rating of essays (Buhrke & Yanico, 1982) found no overall differences in evaluations of male and female writers. Interestingly, when the researchers examined the response of subjects scoring high on a social desirability scale, both male and female subjects assigned higher ratings to female writers.

Regardless of the explanation, the results have implications for the appropriate roles of laboratory and field research. As others have noted (e.g., Berkowitz & Donnerstein, 1982; Mook, 1983), the true value of laboratory research is the explicit examination of well-articulated theoretical propositions and not the estimation of the magnitude of relationships in a population. A good example of the appropriate use of laboratory settings is Schmitt and Lappin's (1980) test of the hypothesis that raters exhibit less confidence and hence have less variability in their ratings of different race or sex ratees. In other words, the theoretical processes or conditions underlying race effects can legitimately be studied in laboratory settings, for these processes may be generalizable to the real world; effect sizes cannot.

The meta-analysis also shows that saliency of blacks in the workgroup may be a moderator. Race effects decline as the percentage of blacks increase. For example, although blacks made up only 22% of the total sample, the percentage of blacks in the 10 studies with the smallest (near zero) effect sizes was 39%. This finding is consistent with theory and research in the social cognition literature. Taylor (1981), for example, has theorized that as the percentage of group members possessing a certain characteristic (race, sex, age, etc.) decreases, that characteristic becomes more salient for processing and recalling information about those particular members. As a result, ratings of those members grow more extreme. The weakness of the present meta-analysis is that the percentage of blacks rated by individual raters was unknown in studies employing multiple workgroups or raters. Moreover, a low percentage does not necessarily imply high saliency. Nonetheless, the results do corroborate predictions from other areas of psychology with real-world data and suggest a fruitful line of research for I/O psychologists.

As noted in the introduction, the results of the present meta-analysis do not directly isolate the effects of racial bias or performance differences in performance evaluations. A closer inspection of the results, though, provides evidence that some portion of rating variance is probably attributable to rater bias. First, raters evaluated same-race ratees higher than different-race ratees. Because the two sets of raters evaluated many of the same ratees, a logical conclusion is that the ratings were biased to some degree. Second, the percentage of blacks in the sample was found to be inversely related to the size of the race effects. This finding suggests that the saliency of a ratee characteristic may be directly related to the degree to which that characteristic is incorporated in performance ratings.

On the other hand, the existence of bias

effects does not preclude the possibility that actual performance differences between races existed. In fact, the data imply that true differences were present in some cases. First, the ratee effect size was quite robust under conditions that focused the attention of the rater on actual job behaviors (i.e., behaviorally based rating scales, rather training). From this perspective, the fact that the conditions did not minimize (and tended to increase) the race effect suggests actual performance differences. Second, the effect size was much larger in field than laboratory settings. Although laboratory investigations typically equate subgroup performance, equal subgroup performance distributions in field settings cannot be assumed. Research on selection tests have shown that white applicants tend to score higher on a variety of preemployment ability measures (e.g., see Schmidt, & Hunter, 1981). Therefore, the higher effect sizes for field studies (with white raters) may, in part, be a reflection of the differences found on the predictor measures.

Consequently, no firm conclusions can be reached regarding the extent to which the results found are due to rater bias or ratee performance. Rather, the evidence suggests that differential ratings are more likely due to some combination of bias and performance differences.

More important than isolating the source of the observed differences between races in this meta-analysis is a consideration of the impact of these differences on future research. Landy and Farr (1980), addressing the source of bias, have conceptualized bias as the application of different mental processes of the rater as a function of ratee subgroup. If one views the racial differences found as due to bias, research efforts should focus on the extent to which (a) race is used as a relevant category for observing, storing, and recalling ratee information; and (b) raters differentially weight job- and non-job-relevant factors in evaluating performance of black and white ratees. The social cognition literature provides directions for examining differences in the way in which information is processed by raters. For example, research on tokenism (Kanter, 1977) might help explain why blacks tend to receive more negative ratings when they constitute only a small percentage of the workforce. In terms of differential weighting of factors, Bass and Turner (1973), Campbell et al. (1973) and others have provided evidence that ratings of black ratees are more related to objective performance measures than white ratees. An interesting follow-up to the present study would be to examine relations between objective and subjective criteria by ratee race using meta-analytic techniques.

The viewpoint that the race effects found in the present study may reflect actual performance differences leads to a research strategy that examines factors in society and organizational settings that foster performance differences. For example, cultural anthropologists such as Sanday (1976) suggest that differences in performance might occur over time because the nondominant culture (e.g., black culture) may not share the common values, perceptions, and institutions of the dominant mainstream culture that defines effectiveness. From an organizational perspective, Ilgen and Youtz (1984) have suggested that lost opportunity factors such as the lack of mentors for blacks, ingroup/outgroup effects (Dansereau, Graen, & Haga, 1975), and self-limiting behaviors perpetuate performance differences between black and white group members.

Consequently, the results of the present study are provocative in directing future research. One research direction is to examine the information processing strategy of the rater, whereas a second direction is to examine contextual factors that might have an impact on individual effectiveness. Based on the present study, both directions seem reasonable and need to be pursued.

In conclusion, this article makes two primary contributions to understanding the impact of race on performance ratings. First, a population effect size has been determined, and this value is the best estimate of what a personnel researcher may expect when collecting ratings. There is little need for future studies that focus exclusively on determining ratee race effect size. Second, the results (particularly of the moderators) should serve to shift the emphasis from examining race effects to a greater understanding of the process underlying race effects. Examination of mean difference in ratings by ratee racial

group does little to increase our understanding of the complex nature of performance differences and racial bias. Instead, we need to develop methods to detect the presence of bias and to investigate the boundaries and covariates of bias in performance evaluations.

References

Arnold, H. J. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior and Human Performance, 29,* 143–174.

Bass, A. R., & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance, *Journal of Applied Psychology, 57,* 101–109.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist, 37,* 245–257.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63,* 301–308.

Bernardin, H. J., & Pence, E. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65,* 60–66.

Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology, 61,* 80–84.

Borman, W. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64,* 410–421.

Brugnoli, G. A., Campion, J. E., & Basen, J. A. (1979). Racial bias in the use of work samples for personnel selection. *Journal of Applied Psychology, 64,* 119–123.

Buhrke, R. A., & Yanico, B. J. (1982, August). *Effects of sex-role ideology on sex bias in performance appraisal.* Paper presented at the meeting of the American Psychological Association, Washington, DC.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in the prediction of job performance: A six year study (Report PR-73-27).* Princeton, NJ: Educational Testing Service.

Cascio, W. F., & Valenzi, E. R. (1978). Relations among criteria of police performance. *Journal of Applied Psychology, 63,* 22–28.

Dansereau, F., Graen, G., & Haga, W. J. (1975). A vertical dyadic linkage approach to leadership within formal organizations. *Organizational Behavior and Human Performance, 13,* 46–78.

deJung, J. E., & Kaplan, H. (1962). Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. *Journal of Applied Psychology, 46,* 370–374.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. (1971). Ethnic group membership as a moderator of the prediction of job performance. *Personnel Psychology, 24,* 609–636.

Feldman, J. M. (1981) Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66,* 127–148.

Fox, H., & Lefkowitz, J. (1974). Differential validity: Ethnic group as a moderator in predicting job performance. *Personnel Psychology, 27,* 209–223.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 10,* 3–8.

Guion, R. M. (1965). *Personnel testing.* New York: McGraw-Hill.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology, 59,* 705–711.

Hays, W. L. (1973). *Statistics for the social sciences.* New York: Holt, Rinehart, & Winston.

Hunter, J. E., Schmidt, F. L., & Jackson, G. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Ilgen, D. R., & Youtz, M. (1984, February). *Factors affecting the evaluation and development of minorities in organizations.* Presented at the Office of Naval Research Conference on minorities entering high-technology careers, Pensacola, FL.

Ivancevich, J. (1979). Longitudinal study of the effects of rater training on psychometric errors in ratings. *Journal of Applied Psychology, 64,* 502–508.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology, 33,* 595–640.

Jones, E. E., & Sigall, H. (1971). The Bogus Pipeline: A new paradigm for measuring affect and attitudes. *Psychological Bulletin, 76,* 349–364.

Kanter, R. M. (1977). *Men and women of the corporation.* New York: Basic Books.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in multidimensional forced choice performance evaluation scale. *Journal of Applied Psychology, 65,* 507–516.

Landy, F. J., & Farr, J. L. (1973). *Police performance appraisal.* University Park, PA: Department of Psychology, The Pennsylvania State University.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107.

Mobley, W. H. (1982). Supervisor and employee race and sex effects on performance appraisals: A field test of adverse impact and generalizability. *Academy of Management Journal, 25,* 598–606.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

Rosenthal, R., & Rubin, D. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504.

Sanday, P. R. (1976). Cultural and structural pluralism in the United States. In P. R. Sanday (Ed.), *Anthropology and the public interest: Fieldwork and theory* (pp. 53–73). New York: Academic Press.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology, 66,* 166–185.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Progress in validity generalization: Comments on Callender and Osburn and further developments. *Journal of Applied Psychology, 67,* 835–845.

Schmidt, F. L., & Johnson, R. H. (1973). Effect of race on peer ratings in an industrial setting. *Journal of Applied Psychology, 57,* 237–241.

Schmitt, N., & Hill, T. E. (1977). Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. *Journal of Applied Psychology, 62,* 261–264.

Schmitt, N., & Lappin, M. (1980). Race and sex as determinants of the mean and variance of performance ratings. *Journal of Applied Psychology, 65,* 428–435.

Sharon, A. T., & Bartlett, C. J. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology, 23,* 251–263.

Taylor, E. L., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology, 4,* 39–47.

Taylor, S. E. (1981). A categorization approach to stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 83–114). Hillsdale, NJ: Erlbaum.

Terborg, J. R., Lee, T. W., Smith, F. J., Davis, G. A., & Turbin, M. S. (1982). Extension of the Schmidt and Hunter validity generalization procedure to the prediction of absenteeism behavior from knowledge of job satisfaction and organizational commitment. *Journal of Applied Psychology, 67,* 440–449.

Toole, D. L., Gavin, J. F., Murdy, L. B., & Sells, S. B. (1972). The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology, 25,* 661–672.

Warmke, D. L., & Billings, R. S. (1979). Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. *Journal of Applied Psychology, 64,* 124–131.

Wendelken, D. J., & Inn, A. (1981). Nonperformance influences on performance evaluations: A laboratory phenomenon? *Journal of Applied Psychology, 66,* 149–158.

Wherry, R. J., Sr., & Bartlett, C. J. (1982). The control of bias in ratings. *Personnel Psychology, 35,* 521–551.

Appendix

Studies Included in the Meta-Analysis of Race Effects

Arnold, B. C. (1968). Comparison of Caucasian and Negro subgroups on criterion indices of overall job effectiveness. *Dissertation Abstracts International, 30*(2), 881B. (University Microfilms No. 69-12, 499).

Arvey, R. D., & Gordon, M. E. (1977). *Final report: Validation of test instruments for the identification of successful and unsuccessful police patrol officers.* Knoxville, TN: The University of Tennessee.

Baehr, M., Saunders, D. R., Froemel, E. C., & Furcon, J. E. (1971). The prediction of performance for black and white police patrolmen. *Professional Psychology, 2,* 46–58.

Bartlett, C. J., Goldstein, I. L., Mosier, S., Hannan, R., Buxton, V., Simmons, V., Cooper, C. (1977). *An analysis of the validity of the PPA police examination for entry-level selection in the Prince George's Police Department.* College Park, MD: Training and Education Research Programs.

Bass, A. R., & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology, 57,* 101–109.

Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology, 61,* 80–84.

Blumberg, M., Farr, J., Landy, F., Neidig, R., Saal, F., & Whitaker, L. (1974). *Report on Standard Pressed Steel validation project.* University Park, PA: The Pennsylvania State University (Department of Psychology).

Brugnoli, G. A., Campion, J. E., & Basen, J. A. (1979). Racial bias in the use of work samples for personnel selection. *Journal of Applied Psychology, 64,* 119–123.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in the prediction of job performance: A six-year study (Report PR-73-27).* Princeton, NJ: Educational Testing Service.

Cascio, W. F., & Valenzi, E. R. (1978). Relationships among criteria of police performance. *Journal of Applied Psychology, 63,* 22–28.

Cox, J. A., & Krumboltz, J. D. (1958). Racial bias in peer ratings of basic airmen. *Sociometry, 21,* 292–299.

deJung, J. F., & Kaplan, H. (1962). Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. *Journal of Applied Psychology, 46,* 370–374.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. (1971). Ethnic group membership as a moderator of the prediction of job performance. *Personnel Psychology, 24,* 609–636.

Farr, J. L., O'Leary, B. S., Pfeiffer, C. M., Goldstein, I. L., & Bartlett, C. J. (1971). *Ethnic group membership as a moderator in the prediction of job performance: An examination of some less traditional predictors.* Pittsburgh: American Institute for Research.

Feldman, J. M., & Hilterman, R. J. (1977). Sources of bias in performance evaluation: Two experiments. *International Journal of Intercultural Relations, 1,* 35–37.

Field, H. S., Bayley, G. A., & Bayley, S. M. (1977). Employment test validation for minority and nonminority production workers. *Personnel Psychology, 30,* 37–46.

Ford, K. A. (1975). Ethnic group differences in employment test/job performance relationships. *Dissertation Abstracts International, 37*(9), 739B. (University Microfilms No. 77-24, 829).

Fox, H., & Lefkowitz, J. (1974). Differential validity: Ethnic group as a moderator in predicting job performance. *Personnel Psychology, 27,* 209–223.

French, J. (1973). *Development and concurrent validation of selection devices for typist clerks: An interjurisdictional research study.* Lansing, MI: Michigan Department of Civil Service.

Gael, S., & Grant, D. L. (1972). Employment test validation for minority and nonminority telephone company service representatives. *Journal of Applied Psychology, 56,* 135–139.

Hakel, M. D., Appelbaum, L., Lyness, K. S., & Moses, J. L. (1982). *Reliable and impartial ratings of management potential.* Unpublished manuscript.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology, 59,* 705–711.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of white and black females. *Personnel Psychology, 29,* 13–30.

Ivancevich, J. M., & McMahon, J. T. (1977). Black-white differences in a goal-setting program. *Organizational Behavior and Human Performance, 20,* 287–300.

Kesselman, G. A., & Lopez, F. E. (1979). The impact of job analysis on employment test validation for minority and nonminority accounting personnel. *Personnel Psychology, 32,* 91–108.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. (1968). *Testing and fair employment: Fairness and validity of personnel tests for different ethnic groups.* New York: New York University Press.

Kniesner, C. (1971). *Report on the validity of our paraprofessional exam series.* Columbus, OH: State of Ohio Personnel Department.

Kraiger, K. (1981). *Measuring police officer performance: Criterion development for the Columbus Police Officer Selection Validation Project.* Columbus, OH.

Kriska, S. D., Hines, C. U., & Katko, D. P. (1983). *Criterion-related validity study of the Columbus, Ohio firefighter job.* Columbus, OH.

Landy, F. J., & Farr, J. L. (1973). *Police performance appraisal.* University Park, PA: The Pennsylvania State University (Department of Psychology).

Lopez, F. M. (1966). Current problems in test performance of applicants: I. *Personnel Psychology, 19,* 10–18.

McCann, F. E., Zupkis, R., Howeth, W. F., & Nichols, G. V. (1976). *The validation of the McCann Associates ESV police officer written test.* Huntington Valley, PA: McCann Associates.

Miller, N., & Angelo, T. (1976). *Concurrent validation study: Income Maintenance Worker I & II.* Columbus, OH: Ohio Department of Administrative Services.

Mobley, W. H. (1982). Supervisor and employee race and sex effects of performance appraisals: A field study of adverse impact and generalizability. *Academy of Management Journal, 25,* 598–606.

Moore, M. H. (1973). An investigation of the influence of ethnic group membership on job attitudes and the relationship between these attitudes and job performance. *Dissertation Abstracts International, 34*(4), 1783B–1784B. (University Microfilms No. 73-22, 922).

Morstain, B. R., & Hsu, L. (1977). *Delaware cooperative police selection study.* Newark, DE: College of Urban Affairs and Public Policy, University of Delaware.

Neidt, C. O. (1968). *Report on differential predictive validity of specified selection techniques within designated subgroups of applicants for Civil Service positions.* Colorado State University: Colorado Civil Rights Commission.

Richards, S. A., & Jaffee, C. L. (1972). Blacks supervising whites: A study of interracial difficulties in working together in a simulated organization. *Journal of Applied Psychology, 56,* 234–240.

Rosenfeld, M., & Thornton, R. F. (1976). *The development and validation of a firefighter selection examination for the City of Philadelphia.* Princeton, NJ: Educational Testing Service.

Rosenfeld, M., & Thornton, R. F. (1976). *The development and validation of a multijurisdictional police officer examination.* Princeton, NJ: Educational Testing Service.

Rosenfeld, M., & Thornton, R. F. (1979). *The development and validation of a police selection examination for the City of Philadelphia.* Princeton, NJ: Educational Testing Service.

Rotter, N., & Rotter, G. S. (1969). *Race, work performance, and merit ratings: An experimental evaluation.* Paper presented at the Eastern Psychological Association Convention at Philadelphia, PA.

Schmidt, F. L., & Johnson, R. H. (1973). Effect of race on peer ratings in an industrial setting. *Journal of Applied Psychology, 57,* 237–241.

Schmitt, N., Merritt, R., Fitzgerald, M., Noe, R. (1981). *Results of the NASSP Assessment Center* (Research Report No. 81-4). East Lansing, MI: Michigan State University.

Schmitt, N., & Lappin, M. (1980). Race and sex as determinants of the mean and variance of performance ratings. *Journal of Applied Psychology, 65,* 428–435.

Selection Consulting Center. (1974). *The validation of entry-level firefighter examinations in the states of California and Nevada.* Sacramento, CA: Selection Consulting Center.

Tenopyr, M. L. (1967). *Race and socioeconomic status as moderators in predicting machine-shop training success.* Presented at the 75th Annual Convention of the American Psychological Association, Washington, DC.

Toole, D. L., Gavin, J. F., Murdy, L. B., & Sells, S. B. (1972). The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology, 55,* 661–672.

Vance, R. J., & Kraiger, K. (1982). *Research manuscript in progress.* Columbus, OH: The Ohio State University.

Received September 21, 1983
Revision received May 14, 1984 ■