DOCUMENT RESUME

ED 255 836 CG 018 168

AUTHOR Ford, J. Kevin; And Others
TITLE The Study of Race Effects in Objective Indices and Subjective Evaluations of Performance: A Meta-Analysis of Performance Criteria.
PUB DATE Mar 85
NOTE 39p.; Portions of this paper were presented at the Annual Convention of the American Psychological Association (92nd, Toronto, Ontario, Canada, August 24-28, 1984).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Blacks; Cognitive Measurement; Criterion Referenced Tests; *Effect Size; Employees; *Job Performance; Meta Analysis; Objective Tests; *Personnel Evaluation; *Racial Factors; *Test Bias; Test Validity; Whites
IDENTIFIERS Absenteeism (Employee); Subjective Tests

ABSTRACT

Although the criterion problem has been acknowledged as critical in personnel research, few attempts have been made to systematically examine the nature and covariates of criterion measures of performance. The present research used meta-analytic techniques to examine the race effect size for objective measures of performance and to compare the relationship between effect sizes for objective indices and subjective ratings. Fifty-three studies were located that included at least one objective index of actual performance, absenteeism or cognitive test performance and one subjective measure of performance for the same group of black and white employees. The corrected average effect sizes across the 53 studies were relatively low but quite similar for the objective and subjective criteria. Moderating effects for the objective criteria were found as race effects were much higher for cognitive than for performance criteria. Subjective ratings had a lower effect size than objective cognitive test scores but were higher than comparable objective performance indices. The implications of the results for personnel research practices were discussed and the need for a better understanding of the constructs underlying criterion measures were emphasized. (Author)

The Study of Race Effects in Objective Indices and Subjective Evaluations of Performance: A Meta-Analysis of Performance Criteria

J. Kevin Ford

Michigan State University

Kurt Kraiger

University of Colorado, Denver

and

Susan L. Schechtman

Michigan State University

March, 1985

Running head: A STUDY OF RACE EFFECTS

Abstract

Although the criterion problem has been acknowledged as critical in personnel research, few attempts have been made to systematically examine the nature and covariates of criterion measures of performance. The present research used meta-analytic techniques to examine the race effect size for objective measures of performance and to compare the relationship between effect sizes for objective indices and subjective ratings. Fifty-three studies were located that included at least one objective index of actual performance, absenteeism or cognitive test performance and one subjective measure of performance for the same group of black and white employees. The corrected average effect sizes across the 53 studies were relatively low but quite similar for the objective and subjective criteria. Moderating effects for the objective criteria were found as race effects were much higher for cognitive than for performance criteria. Subjective ratings had a lower effect size than objective cognitive test scores but were higher than comparable objective performance indices. The implications of the results for personnel research practices were discussed and the need for a better understanding of the constructs underlying criterion measures was emphasized.

The Study of Race Effects in Objective Indices and Subjective Evaluations of Performance: A Meta-Analysis of Performance Criteria

A continual issue of concern for organizations is racial discrimination in personnel practices. Considerable research has revolved around hiring practices, issues of test validity and differential prediction for minority and nonminority applicants. The research indicates that the average test score tends to be lower for minorities but that predictor tests appear equally valid for both minority and majority group members (e.g., Bartlett, Bobko, Mosier, & Hannan, 1978). These results have led researchers to conclude that predictors such as cognitive ability tests are fair to minority applicants as they do not systematically underestimate the expected job performance of minority groups (Schmidt & Hunter, 1981).

The research on personnel testing has mainly focused on predictor rather than criterion related issues. An implicit assumption underlying this focus is that the criteria employed in test validity and fairness studies are job relevant and unbiased. This neglect of criterion issues is surprising given the long standing concern of personnel and other applied psychologists over the "criterion problem" (Cascio, 1982; Smith, 1976; Wallace, 1965). The quality of criterion related validity and test fairness studies is heavily dependent upon the appropriateness of the criterion used (Burke, 1984; Scott & Hamner, 1975). Additionally, testing standards

require researchers to investigate the relevance of the criterion and to study the possibility that irrelevant factors may cause criterion bias (American Psychological Association, Division of Industrial and Organizational Psychology, 1980). The purpose of the present study is to identify the criteria used in testing studies and to investigate the extent to which race effects are present in various types of performance criteria.

Performance ratings are the most often used criterion in validation studies (Landy & Farr, 1980; Schmitt, Gooding, Noe, & Kirsch, 1984). Despite their widespread use, ratings have been criticized as highly vulnerable to rater biases such as halo, leniency and stereotyping. This vulnerability to intentional or inadvertent racial bias has led researchers to question the usefulness of ratings as criteria in test fairness studies. For example, in a critique of the Educational Testing Service Project on racial bias, (Campbell, Crooks, Mahoney & Rock, 1973), Anastasi (1973) questioned the relevancy of ratings as criterion measures for test validation when different ethnic groups were involved. In another critique of the project, Wallace (1973) cited the lack of relevance (low intercorrelations with work samples), rater bias (a rater by ratee interaction) and the spurious nature of rating reliability estimates as providing the final "stake" for the internment of supervisory ratings as test validation criteria.

A recent meta-analysis of race effects in ratings by Kraiger and Ford (1985) provided a direct examination of the relationship

between subjective ratings and race. The study revealed a relatively low but consistent ratee race effect size and a rater by ratee race interaction. Whites rated white ratees higher than black ratees and black raters assigned higher ratings to blacks than to whites. Moderator analyses of the ratings of white raters revealed that rating scale format (behavior based/trait), rater training (offered/not offered) and rating purpose (administrative/ research) had minimal impacts on the size of the race effects found. Race effects, though, were more likely to be found in field than laboratory settings and the effect size was higher (favoring white ratees) when black ratees constituted a smaller percentage of the workforce.

The Kraiger and Ford (1985) study was limited to the examination of race effects in subjective ratings and therefore could not directly isolate the relative contributions of ratee performance and rater biases to the rating differences found. The interaction effect found for race of rater and ratee suggests that some degree of bias is present in the ratings as both white and black raters evaluated many of the same ratees. Nevertheless, the effects found do not preclude the possibility that actual performance differences between races exist. Albright (1973), for example, has argued that it is premature to dismiss racial differences in ratings as due largely to bias without a comparison to more objective criterion measures such as turnover and productivity criteria which more closely reflect real life decisions and actions in organizations.

Although the distinction between subjective and objective criteria is problematic and somwhat arbitrary (Guion, 1981; Muckler, 1982), a number of studies have examined the relationship of performance ratings and objective measures of performance. Laboratory and simulation studies have generally found strong relationships between actual performance and performance ratings (Bigoness, 1976; Borman, 1978; Schmitt & Lappin, 1980). The results from field studies are more conflicting; some studies have found low relationships between objective indices and subjective ratings of job performance (Alexander & Wilkins, 1982; Hausman & Strupp, 1955; Seashore, Indik & Georgopolous, 1960) while other studies have found more substantial relationships (Bass & Turner, 1973; Kirchner, 1960).

Two recent meta-analyses support the findings of low to moderate relationships between subjective and objective measures. Heneman (1983) examined the relationship of cost or profit related criteria and overall effectiveness ratings across fourteen studies and found a mean corrected correlation of .28 and a large 95% confidence interval which included zero. Hunter (1983) viewing ratings as the dependent variable, attempted to model the relationships among ability tests, job knowledge tests, work samples and performance ratings. The resultant multiple correlation for the prediction of the supervisory ratings from these objective sources of information was .42.

While the results of the above correlational studies indicate that ratings and objective measures are related, there remains a

large amount of variance unaccounted for in both sets of measures. Guion (1983) has suggested that the impact of exogenous variables, such as ratee and rater characteristics and contextual factors, must be included in any model that attempts to increase our understanding of the relationships among objective and subjective criteria. Similarly, Scott and Hamner (1975) and Mobley (1982) have stressed the need for more research on objective and subjective measures of performance and their relaitonship to possible contaminating factors such as race and sex.

Researchers have continued to call for increased use of objective measures of performance in personnel research, especially measures that have utility value to the organization (Tenopyr & Oeltjen, 1980; Zedeck & Cascio, 1984). An implicit and often untested assumption underlying the use of objective measures is that they are less prone to biases than subjective measures. This orientation is demonstrated in the studies above which examined the relevancy of ratings as a function of their relationship to objective performance indices. While less prone to certain biases inherent in more judgmental measures, objective measures are contaminated to an unknown degree (Cascio & Valenzi, 1978). Unlike the extensive investigation of biasing influences in subjective ratings, there have been few systematic attempts to investigate the nature and covariates of objective criterion measures. In particular, research is lacking which examines the relationships between employee race and various types of objective criterion measures.

The present study builds upon the work of Kraiger and Ford (1985) and examines the relationship of race and criterion measures through meta-analytic procedures. While the interpretation of racial differences is problematic without some ultimate measure of performance, differences in effect size may be more readily interpretable as bias or relevance with multiple criterion measures. Therefore, the two major goals of the study are to: (1) investigate differences in race effect size among different types of objective criteria; and (2) directly compare the relationship between effect sizes for objective indices and subjective ratings of performance.

Method

An attempt was made to locate, summarize and analyze the results of all published studies and a number of unpublished studies reporting at least one objective index and one subjective rating of performance for the same sample of black and white employees. A majority of the studies were used in a previous analysis of race effects in performance ratings (Kraiger & Ford, 1985). Additional studies were located by systematically reviewing the recent literature and by soliciting responses from researchers active in test validation and performance assessment. In some cases more than one sample possessing the above characteristics was described in the same report or article. As a result, a total of 53 samples (25 published and 28 unpublished) were located. A complete list of studies is presented in the appendix.

Analysis

The meta-analysis cumulated point-biserial correlations between ratee race (arbitrarily coded White = 1, Black = 0) and objective indices of performance and subjective ratings in order to compute mean effect sizes ($\underline{r}_{pb}$) and variances (σ^2_{pb}) across studies. Point-biserial correlations were typically calculated from either a t-test of group differences or reported group means and standard deviations. An estimate of variance due to sampling error (σ^2_e) and the population variance for effect sizes (σ^2_ρ) were computed for both subjective and objective criteria using procedures explained by Hunter, Schmidt, and Jackson (1982). The estimated standard error (σ_e) was used to establish confidence intervals around the appropriate $\underline{r}_{pb}$ to test the hypothesis that $\underline{r}_{pb} = 0$ in the population.

Since 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 prior to cumulation. Estimated sampling error was then adjusted for this correction.

Coding of study characteristics and effect size calculations for the meta-analysis were completed by two of the authors and differences resolved through consensus of the three authors or recalculation of the appropriate statistic.

Moderator Analyses

The population variance (σ^2_ρ) estimates actual study-to-study variation in effect sizes with variance due to small samples and unequal sample sizes removed. Hunter et al. (1982) stated that this corrected variance may be trivial and due to statistical artifacts or may be nontrivial and suggest possible moderators. The two tests of triviality of the corrected variances suggested by Hunter et al. (1982) revealed significant chi squares, $\chi^2(53, \underline{N} = 10,222) = 434.97$, $\rho < .01$, for the objective data and $\chi^2(53, \underline{N} = 9,443) = 167.22$, $\rho < .01$, for the subjective data, and a small ratio of sampling error to true variance ratio for the sample of objective (.13) and subjective (.33) criteria. These results suggested that the effects were non-trivial and supported the investigation of potential moderators in the data.

A total of 44 different objective criteria were used across the 53 samples. To compare effect sizes across criterion types, it was necessary to reduce this set of criteria to a smaller, more conceptually meaningful number of categories. Since an adequate classification of objective criteria was not found, a categorization system was developed.[1] For this task, seven advanced graduate students free sorted the 44 criteria into categories. Five stable categories of criteria emerged from the sorting and were labeled as training tests, job knowledge tests, absenteeism and tardiness, direct performance (e.g., units produced, shortages) and indirect performance (e.g., accidents, customer complaints). Five additional

graduate students independently resorted the 44 criteria into the five derived categories. The high level of agreement (91%) resulting from the sorting task demonstrated the reliability of the criterion categories. The researchers examined the disagreements and came to a consensus as to the appropriate categorization of the criteria.

The five categories were further reduced to three by combining the indirect and direct performance measures into a performance indices category and by combining the training and job knowledge tests into a cognitive criteria category. To maintain independence of observation, multiple objective criteria belonging to a particular category were combined if they were related to the same subjective rating. This combination of criteria reduced the sample size for the moderator analysis from 53 to 49. The three categories of performance indicators ($\underline{N}$ = 20), absenteeism ($\underline{N}$ = 13) and cognitive criteria ($\underline{N}$ = 16) provided the small number of categories needed to meaningfully examine differences in race effect size by criterion category. It should be noted that analyses conducted with the five criterion categories yielded similar results as those to be presented for the three criterion categories.

For the subjective evaluations, overall ratings of effectiveness were available from each of the 53 samples. Ten samples also included a specific rating that matched the type of objective criteria gathered in that study (e.g., a rating of job knowledge and a test of job knowledge). When available, the specific ratings

were used in place of the overall effectiveness ratings in the analyses to more closely match the objective criterion.[2] Each subjective rat ng was matched by category with the objective criterion in the same sample and average effect sizes across samples were calculated. This allowed for the direct comparison of objective and subjective effect sizes.

It should be noted that nearly all raters were white so that effect sizes for ratings by black raters could not be compared to objective data. In addition, few studies reported criterion reliability data to use in correcting for attenuation. The limited literature relevant to criterion reliability indicates that performance (except for repetitive jobs) (Rothe, 1947; Rambo, Chomiak & Price, 1983) and absenteeism measures (Muchinsky, 1977) are particularly unstable. Based on this literature, the reliabilities for objective measures of performance and absenteeism were set at a conservative level of .60. The reliability for cognitive tests was set at .80 (Hunter & Hunter, 1984) while the reliability for the performance ratings was estimated to be .70 (Kraiger & Ford, 1985).

Finally, moderating effects were shown by classifying the studies into relevant subsamples and recomputing subsample $\underline{r}_{pb}$'s, σ's and confidence intervals. Differences in subgroup effect sizes were tested for significance by a procedure adapted from Rosenthal and Rubin (1982) and previously used in the meta-analysis of race effects (Kraiger & Ford, 1985). Rosenthal and Rubin (1982) have shown that their derived quotient is distributed as the standard

normal deviate, Z. A significant Z indicates that effect sizes differ between at least two moderator subgroups.

Correlation of Effect Sizes

To further examine the relationship between objective and subjective criteria, the effect size for the objective measure and the subjective measure for each sample were correlated. A correlation of effect sizes was computed for the 53 samples overall and within the three criterion categories of performance, absenteeism and cognitive criteria. The results provide an indication of the covariation of race effect size between subjective and objective measures (i.e., the extent to which a sample with a large race effect size on a subjective rating tended to have a large race effect size on the objective criterion of performance).

Results

The results of the meta-analysis are presented in Table 1. The table shows sample sizes for whites and blacks, total sample sizes, corrected effect sizes, variance estimates, and confidence intervals for the 53 studies with an objective and subjective measure of performance. Information is also presented for the three criterion categories of performance indicators, absenteeism and cognitive criteria.

Insert Table 1 about here

The best estimate of the population effect size is the mean point-biserial correlation corrected for unreliability. For the

objective indices, this estimate was .209 based on a sample of 10,222 (7,405 whites and 2,817 blacks) employees. For the subjective ratings, the mean point-biserial correlation was quite similar (.204) and was based on a total sample size of 9,443 (6,791 whites and 2,652 blacks). The 95% confidence interval for the objective indices ($.06 < \rho < .36$) and the subjective ratings ($.04 < \rho < .37$) both excluded zero. This finding indicates that whites are rated higher than black ratees and that the level of performance for whites is higher on the objective performance indices.

Moderator Analyses

Table 2 presents the results of the moderator analyses which compared the race effect size found across the three criterion categories. The table shows the corrected mean point-biserial correlations for objective indices and subjective ratings across the three criterion categories. This allowed for the testing of differences in effect size for the three types of objective measures and for the differences between objective and subjective measures of performance. The test of differences in effect size across the three criterion categories for the objective and subjective indices of performance are provided in the rows of the table. The comparisons between objective and subjective effect sizes within each criterion category (i.e., performance indicators, absenteeism, cognitive criteria) are presented in the three columns in Table 2.

Insert Table 2 about here

The results of the Z tests indicated that there were significant differences in effect sizes across the three types of objective indices ($z = 5.63$; $\rho < .01$). Inspection of the three average effect sizes revealed that the difference resulted from the larger effect size for the cognitive criteria than for either the performance or absenteeism criteria. Differences in effect sizes for the subjective ratings across the three criterion categories were nonsignificant ($z = 0.45$; $\rho > .05$).

Within criterion categories, Z tests were calculated for the comparison of objective and subjective criteria. In the performance criterion category, the effect size for the objective performance indicators was significantly smaller than the comparable subjective rating effect size ($z = 3.51$; $\rho < .01$). Conversely, the effect size for the objective cognitive criteria was significantly larger than the effect size for the subjective ratings ($z = 2.13$; $\rho < .01$). There was no significant difference in the objective and subjective rating effect sizes for the absenteeism category.

Correlation of Effect Size

For the overall analysis of 53 samples, the correlation of objective and subjective effect size was .43 ($\rho < .01$). An examination of the three criterion categories revealed significant correlations between objective and subjective effect size for the performance indicators ($N = 20$, $r = .44$; $\rho < .05$) and cognitive

criteria (N = 16; r = 55; $\rho < .05$). The correlation of the two effect sizes for the absenteeism data was nonsignificant (N = 13; r = .17; $\rho > .05$).

Discussion

One goal of the present meta-analysis was to document the extent of racial differences on objective criteria. The overall results showed a relatively small but significant race effect size for objective criteria. The investigation of differences across types of objective criteria found that the average effect size for the cognitive (training and job knowledge tests) criteria was larger than the effect size for absenteeism and performance data. The mean effect size was only slightly higher for the performance than for the absenteeism data.

The second goal of the study was to compare effect sizes for subjective rating criteria and the objective measures. The results indicated that across all studies and criterion categories, the effect size for subjective and objective criteria were virtually identical. White employees were rated higher and were performing at a higher level (as measured by the objective criteria) than black employees. Nevertheless, differences in effect size were evident between objective and subjective measures within criterion category level. First, cognitive tests enlarged the differences between races relative to a matched set of subjective ratings. Second, actual performance indicators revealed smaller differences relative to a matched set of subjective ratings. The magnitude of

effect size was similar for the objective and subjective ratings for the absenteeism category.

Closer inspection of the results in relation to current perspectives in personnel research reveals a number of interesting patterns. First, despite evidence from test fairness studies that blacks typically score about one standard deviation (r_{pb} of approximately .50) below whites on pre-hire aptitude tests (Hunter, Schmidt, & Hunter, 1979) and continue to demonstrate lower performance on job knowledge exams (r = .34 in the present study), subgroup differences in actual on-the-job performance (r = .16) do not appear as large. This result implies that both aptitude tests for selection and job knowledge tests measure some construct correlated with race but somewhat irrelevant to actual job performance (Wallace, 1973).

Second, while job knowledge has been found to be strongly related to supervisory ratings (Hunter, 1983), race effects for ratings are smaller than for cognitive criteria such as job knowledge tests. This result implies that although differences in job knowledge may be incorporated in their ratings, raters must use other factors that have the effect of reducing race effects in ratings. Since the effect size for ratings is closer to the effect size found for actual job performance, it could be argued that performance information is another critical factor incorporated into ratings. Interestingly, Ford, Schechtman and Kraiger (1985) found that white raters placed more weight on objective job performance indices but a similar weight to job knowledge information when rating blacks than when rating white ratees.

Third, race effects tended to covary between objective and subjective measures (e.g., studies with large (or small) race effect sizes for ratings tended to have a large (small) effect size for the same sample of employees on the objective measure. These results counter assertions that objective measures may be preferable to ratings as they are less prone to race effects (Boehm, 1972; Bray & Moses, 1972). On the other hand, the relatively high degree of consistency in the effect sizes found across multiple criterion measures suggests that the race effects found in subjective ratings cannot be solely attributed to rater bias.

The results of the meta-analysis have implications for test validity and fairness research as well as for future research on criterion measurement. Schmidt and his colleagues (e.g., Schmidt & Hunter, 1981; Schmidt, Hunter, Pearlman, & Shane, 1979; Schmidt, Pearlman, & Hunter, 1981) have conducted a number of reviews and studies which indicate that validities are similar for different races, predictors are fair to minorities and validities are generalizable across situations. In conducting these studies, criteria (performance ratings, job knowledge tests, job proficiency scores) are argued to be substitutable because the intercorrelation of criterion measures approaches unity when corrected for measurement reliability (Pearlman, Schmidt, & Hunter, 1980).

The results of the present meta-analysis indicate that criterion measures are not substitutable in terms of expected race effect size. For example, the use of a job knowledge test as a criterion of performance will result in a larger expected race effect size than for other possible criterion measures such as actual performance indices or subjective ratings. While ratings have been criticized for bias, the use of cognitive criteria may also result in what has been labeled apparent but false non-discrimination (Cascio, 1982). In this case, the same factors that act to depress the performance of a subgroup on predictor aptitude tests are likely to be present in the job knowledge criteria. Similarly, Burke (1984) has recently suggested that some component of generalized validity reflects a spurious association between predictor and non job-related biases on the criterion.

Therefore, although it is important to know that validity coefficients are similar between races and that regression equations do not underpredict job performance, it is also premature to consider criteria as substitutable. The present meta-analysis showed that different measures of performance have somewhat different relationships to the exogenous variable of race. This result points to the need for a greater understanding of the factors or constructs being measured by criterion measures and their relationships to other (predictor) constructs.

A construct validation approach to criterion measurement requires the cumulative understanding of a construct that comes

from a sizeable body of empirical data; the kind of data collected in research exploring the network of associations and situations in which a measure acts (Nunnally, 1967). The present study has provided an association of one demographic variable of race to objective and subjective measures of performance. Future research is needed which focuses on the job relevant and irrelevant factors underlying these criterion measures. Without such analyses, it is difficult to interpret the reasons for racial differences found on the criterion measures.

For example, differences in actual performance or job knowledge tests may simply reflect the selection policies of organizations. Because of differing selection ratios, organizations may be able to select only high ability applicants from the white applicant pool but must select from a wider diversity of ability for the minority applicants. These differences, at the time of selection, are then reflected on the criterion measures (Kroeck, Barrett, & Alexander, 1983). Actual performance differences may also reflect to some unknown extent organizational practices such as blacks being placed on older equipment, given less desirable work territories or sent into high risk situations in which accidents are more likely to occur. In this case race acts as an indicator of underlying sources of job irrelevant variance. An interesting implication from this perspective is that performance ratings may be highly "relevant" in the sense that they reflect existing organizational conditions and practices rather than the inherent biases of the

rater. Racial performance differences may also reflect individual level factors such as the lack of mentors and self limiting behaviors (Ilgen & Youtz, 1984) or simply lower job tenure (Bernardin, 1984) for blacks than for whites.

Unfortunately, data relevant to these issues were not available for systematic analysis in this sample. Interestingly, O'Connor and his colleagues (O'Connor, Peters, Pooyan, Weekly, Frank, & Erenkrantz, 1984; Peters, O'Connor, & Rudolf, 1980) have been conducting a line of research regarding the impact of organizational constraints on job performance. Such research, while not directly addressing the issues above, provides a useful model for the type of research needed to better understand the constructs we are actually measuring when gathering "objective" measures of performance as well as a better understanding of the constraints affecting the relationship between individual differences and job performance.

To increase our understanding of the constructs underlying subjective ratings, Guion (1983) has suggested the need for longitudinal studies in which records of evaluations, ratee characteristics, any changes in rater characteristics and changes in circumstances over time are examined for cyclical effects either in contextual variables or performance. Another research direction is the investigation of the processing of job relevant and job irrelevant factors by raters in the evaluation of performance. Pettigrew (1979), for example, has suggested the counterintuitive premise that a positivity bias is operating in which ratings

reflect an inflation of the majority group's ratings rather than an assumed deflation of ratings against minorities. Positivity bias argues that majority group members receive higher ratings because compensatory job irrelevant factors (e.g., familiarity with the rater) are considered for majority group members while minorities group ratings are more reflective of true performance levels (Kraiger, 1981). While these behavior patterns have been recognized, research on the weighting of factors has not been systematically applied to the performance evaluation domain. Multidimensional scaling, policy capturing, and information processing boards from decision making research (e.g., Billings & Marcus, 1983) provide useful methodologies for conducting research on the effects of job relevant and irrelevant influences on rater judgments.

The results of the present meta-analysis, when combined with the previous work of Kraiger and Ford (1985), provide a comprehensive analysis of race effects in performance criteria. While the data upon which this analysis is based should be continually updated, research needs to go beyond the simple focus on whether race effects occur in criterion measures. Regardless of the specific research direction taken relevant to criterion measurement, we conclude by paraphrasing Wallace's (1974) advice to observe high validity coefficients from test fairness and validity generalization studies with more suspicion and less euphoria, to seek instruction from them rather than reassurance, to use them to create constructs rather than to build empires and to worry about criteria first and predictors later.

References

Albright, L. E. (1973). Sources of bias in the prediction of job performance: Implications for employers in industry. In L. A. Crooks (Ed.), An investigation of sources of bias in the prediction of job performance. Princeton, NJ: Educational Testing Service.

Alexander, E. R., & Wilkins, R. D. (1982). Performance rating validity: The relationship of objective and subjective measures of performance. Group and Organization Studies, 7, 485-496.

American Psychological Association, Division of Industrial and Organizational Psychology. (1980). Principles for the validation and use of personnel selection procedures (2nd ed.). Berkeley, CA: Author.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 31, 233-241.

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

Bernardin, H. J. (1984, August). An analysis of black-white differences in job performance. Presented at the 44th Annual Meeting of the Academy of Management, Boston, MA.

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.

Billings, R. S., & Marcus, S. A. (1983). Measures of compensatory and noncompensatory models of decision behavior: Process tracing versus policy capturing. Organizational Behavior and Human Performance, 31, 331-352.

Boehm, V. R. (1972). Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. Journal of Applied Psychology, 56, 33-39.

Borman, W. C. (1978). Exploring upper limits on reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Bray, D. W., & Moses, J. L. (1972). Personnel selection. Annual Review of Psychology, 23, 545-576.

Burke, M. J. (1984). Validity generalization: A review and critique of the correlation model. Personnel Psychology, 37, 93-116.

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. (Final Project Report PR-73-37). Princeton, NJ: Educational Testing Service.

Cascio, W. F. (1982). Applied Psychology in Personnel Management (2nd ed.). Reston, VA: Reston.

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

Ford, J. K., Schechtman, S. L., & Kraiger, K. (1985). The relationship among criteria as a function of subgroup membership: An integrative review. Unpublished manuscript.

Guion, R. M. (1983). Comments on Hunter. In F. J. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, NJ: Erlbaum.

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

Hausman, S. E., & Strupp, H. H. (1955). Non-technical factors in supervisors' ratings of job performance. Personnel Psychology, 8, 201-217.

Heneman, R. (1983, March). The relevance of supervisory ratings to measures of more "ultimate" criterion: A meta-analytic investigation. Presented at the Fourth Annual Industrial/ Organizational Psychology and Organizational Behavior Graduate Student Convention, Chicago, IL.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance and supervisor ratings. In F. J. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (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.

Kirchner, W. E. (1960). Predicting ratings of sales success with objective performance information. Journal of Applied Psychology, 44, 398-403.

Kraiger, K. (1981, March). Objectivity as a source of bias in ratings: A case of positivity bias. Presented at the 24th Annual Midwest Academy of Management Conference, Chicago, IL.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. Journal of Applied Psychology, 70, 56-65.

Kroeck, K. G., Barrett, G. V., & Alexander, R. A. (1983). Imposed quotas and personnel selection: A computer simulation study. Journal of Applied Psychology, 68, 123-136.

Landy, F. J., & Farr, J. L. (1980). Performance ratings. Psychological Bulletin, 87, 72-107.

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

Muchinsky, P. M. (1977). Employee absenteeism: A review of the literature. Journal of Vocational Behavior, 10, 316-340.

Muckler, F. A. (1982). Evaluating productivity. In M. D. Dunnette & E. A. Fleishman (Eds.), Human performance and productivity: Vol 1. Human capability assessment. Hillsdale, NJ: Erlbaum.

Nunnally, J. C. (1967). Psychometric Theory. New York: McGraw-Hill.

O'Connor, E. J., Peters, L. H., Pooyan, A., Weekley, J., Frank, B., & Erenkrantz, B. (1984). Situational constraint effects on performance, affective reactions and turnover: A field replication and extention. Journal of Applied Psychology, 69, 663-672.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Peters, L. H., O'Connor, E. J., & Rudolf, C. J. (1980). The behavioral and affective consequences of performance relevant situational variables. Organizational Behavior and Human Performance, 25, 79-96.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. Personality and Social Psychology Bulletin, 5, 461-476.

Rambo, W. W., Chomiak, A. M., & Price, J. M. (1983). Consistency of performance under stable conditions of work. Journal of Applied Psychology, 68, 78-87.

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

Rothe, H. F. (1947). Output rates among machine operators: Distributions and their reliability. Journal of Applied Psychology, 31, 484-489.

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., & Shane, G. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 32, 257-281.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic americans: A review and analysis. Personnel Psychology, 33, 706-721.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

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.

Scott, W. E., & Hamner, W. C. (1975). The influence of variations in performance profiles on the performance evaluation process: An examination of the validity of the criterion. Organizational Behavior and Human Performance, 14, 360-370.

Seashore, S. E., Indik, B. P., & Georgopolous, B. S. (1960). Relationships among criteria of job performance. Journal of Applied Psychology, 44, 195-202.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of the criteria. In M. D. Dunnette (Ed.), The Handbook of Industrial/Organizational Psychology. Chicago, IL: Rand McNally.

Tenopyr, M. L., & Oeltjen, P. D. (1982). Personnel selection and classification. Annual Review of Psychology, 33, 581-618.

Wallace, S. R. (1965). Criteria for what? American Psychologist, 20, 411-417.

Wallace, S. R. (1973). Sources of bias in the prediction of job performance: Implications for future research. In L. A. Crooks (Ed.), An investigation of sources of bias in the prediction of job performance. Princeton, NJ: Educational Testing Service.

Wallace, S. R. (1974). How high the validity? Personnel Psychology, 27, 397-407.

Zedeck, S., & Cascio, W. F. (1984). Psychological issues in personnel decisions. Annual Review of Psychology, 35, 461-518.

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), 818B. (University Microfilms No. 69-12, 499).

Baehr, M. E., Saunders, D. R., Froemel, E. C., & Furcon, J. E. (1971). The prediction of performance for black and for white police patrolmen. Professional Psychology, 2, 46-57.

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 Educational 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.

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 No. 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.

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

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

Fox, H., & Lefkowitz, J. (1974). Differential validity: Ethnic group as a moderator in predicting job performance. Personnel Psychology, 27, 209-233.

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

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.

Kraiger, K. (1981). Measuring police officer performance: Criterion development for the Columbus police officer selection 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.

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

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

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.

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

Author Notes

Portions of this paper were presented at the 92nd Annual Convention of the American Psychological Association, August, 1984, Toronto, Ontario.

Address correspondence to J. Kevin Ford, Department of Psychology, Michigan State University, 129 Psychology Research Building, East Lansing, MI 48824-1117.

The authors express their appreciation to Daniel Ilgen, Neal Schmitt and Mary Zalesny for reviewing an earlier draft of this paper.

Footnotes

[1]For example, previous attempts to categorize criteria have grouped together error counts, attendance, job samples, tenure and job knowledge tests (Hunter, Schmidt, & Hunter, 1979) or grouped work samples, training, and rating criteria (Schmidt, Pearlman, & Hunter, 1980).

[2]The average effect size for the specific ratings and the overall ratings from the same studies was quite similar ($\bar{x}$ = .18, σ^2 = .12; $\bar{x}$ = .12, σ^2 = .19; respectively) and analyses conducted with only overall ratings yielded similar results as those presented for the studies with specific ratings.

Table 1

Mean and Variance of Race Effects by Type of Objective and Subjective Criterion

	No. Studies	Sample size N_W	N_B	N_{Total}	Corrected[a] $\bar{r}_{pb}$
Total Sample					
Objective criteria	53	7405	2817	10222	.209
Subjective criteria	53	6791	2652	9443	.204
Criterion Categories					
Performance indicators					
Objective	20	3260	1027	4287	.159
Subjective	20	3122	1008	4130	.221
Absenteeism					
Objective	13	1529	622	2151	.112
Subjective	13	1563	658	2221	.149
Cognitive					
Objective	16	2371	1018	3389	.336
Subjective	16	1909	873	2782	.226

(table continued)

Table 1 (cont.)

Mean and Variance of Race Effects by Type of Objective and Subjective Criterion

	σ_r^2	σ_e^{2b}	σ_ρ^2	Corrected confidence intervals
Total sample				
Objective criteria	.045	.006	.040	$.058 < \rho < .360$
Subjective criteria	.018	.007	.011	$.037 < \rho < .371$
Criterion categories				
Performance indicators				
Objective	.046	.006	.040	$.008 < \rho < .310$
Subjective	.018	.007	.011	$.058 < \rho < .384$
Absenteeism				
Objective	.017	.008	.009	$-.062 < \rho < .286$
Subjective	.019	.007	.012	$-.014 < \rho < .312$
Cognitive				
Objective	.041	004	.036	$.205 < \rho < .467$
Subjective	.017	.007	.010	$.065 < \rho < .387$

[a]Corrected for unequal sample sizes and for attenuation.

[b]Corrected for added variance due to correction to point-biserial for unequal sample sizes.

Table 2

Moderator Analyses Across Criterion Categories and for the Comparison of Objective and Subjective Criteria

	Criterion category			
Criterion type	Performance $\bar{r}_c$	Absenteeism $\bar{r}_c$	Cognitive $\bar{r}_c$	Z test[b]
Objective indices	.159[a]	.112	.336	5.36*
Subjective ratings	.221	.149	.226	0.45
Z test	3.51*	0.80	2.13*	

[a]Entries are mean point-biserial correlations as corrected for unequal sample sizes and attenuation for unreliability.

[b]Z tests are based on Rosenthal and Rubin (1982).

*$p < .01$