Journal of Applied Psychology
2003, Vol. 88, No. 4, 694–706

0021-9010/03/$12.00 DOI: 10.1037/0021-9010.88.4.694

Ethnic Group Differences in Measures of Job Performance: A New Meta-Analysis

Philip L. Roth
Clemson University

Allen I. Huffcutt
Bradley University

Philip Bobko
Gettysburg College

The authors conducted a new meta-analysis of ethnic group differences in job performance. Given a substantially increased set of data as compared with earlier analyses, the authors were able to conduct analyses of Black–White differences within more homogeneous categories of job performance and to reexamine findings on objective versus subjective measurement. Contrary to one perspective sometimes adopted in the field, objective measures are associated with very similar, if not somewhat larger, standardized ethnic group differences (*d*s) than subjective measures across a variety of indicators. This trend was consistent across quality, quantity, and absenteeism measures. Further, work samples and job knowledge tests are associated with larger *d*s than performance ratings or measures of absenteeism. Analysis of Hispanic–White standardized differences shows that they are generally lower than Black–White differences in several categories.

The issue of majority–minority differences in job performance is an important issue for academics as well as for practitioners and managers. From an academic standpoint, one could suggest that a substantial portion of a selection researcher's role is to predict job performance (Viswesvaran, 2001) and that majority–minority differences in performance are an important part of understanding this issue (Martocchio & Whitener, 1992). From a practitioner or managerial standpoint, it is socially and legally important to hire and maintain a diverse workforce. In addition, the issue of majority–minority differences in performance may be important for job promotion. If performance in a given job is a partial determinate of promotion to a higher level job, differential performance in the feeder job may result in differential promotion rates among ethnic groups.

Previous meta-analytic research has shed some light on these issues. Previous work has generally suggested that measured performance of Whites is, on average, greater than the measured performance of Blacks (J. K. Ford, Kraiger, & Schechtman, 1986; Sackett & Dubois, 1991), but that objective measures of performance often show smaller differences between ethnic groups than do subjective measures (J. K. Ford et al., 1986). This is an important issue because objective measures of performance are thought to be less open to bias than are subjective measures (Schmidt & Hunter, 1981; Rotundo & Sackett, 1999).

Although previous meta-analyses have aided our understanding of ethnic differences in job performance, there is much more to be learned. For example, there appear to be no meta-analyses of Hispanic–White differences. Further, it would appear possible to refine our understanding of Black–White differences, as well as update estimates with new studies. Some previous work has emphasized the analyses from a combination of both laboratory and field studies (Kraiger & Ford, 1985); however, it has been noted that laboratory studies have noticeably different results from the field studies, and this pattern of differences has influenced results that researchers might find in field studies (Sackett & Dubois, 1991). Other meta-analyses have focused on comparisons of objective and subjective measures of performance (J. K. Ford et al., 1986); however, limited sample sizes have forced researchers to combine measures of performance into somewhat heterogeneous categories of performance, despite substantial efforts to the contrary. For example, researchers have had to use a single category of job performance including accidents, performance ratings, and complaints. In further refining our knowledge, it would be useful to consider more homogeneous categories of performance (including the issue of typical vs. maximum measures of performance) to help researchers, practitioners, and managers more accurately understand differences in majority and minority group performance on the job.

The purpose of this article is to meta-analyze majority–minority differences in measures of job performance. We focus on Black–White and Hispanic–White standardized differences because they are two of the largest minority groups in the United States. We do not focus on either Asians or Native Americans because there was not sufficient information available to analyze.

Philip L. Roth, Department of Management, Clemson University; Allen I. Huffcutt, Department of Psychology, Bradley University; Philip Bobko, Department of Management, Gettysburg College.

We thank John Bernardin and J. Kevin Ford for their help. We appreciate their thoughts, assistance in finding studies, and collegial interactions.

Correspondence concerning this article should be addressed to Philip L. Roth, Department of Management, 101 Sirrine Hall, Clemson University, Clemson, South Carolina 29634-1305. E-mail: rothp@clemson.edu

Studies Focusing on Rating Criteria

Several major studies have focused on ratings as a measure of performance. The first major meta-analysis to move this area beyond narrative reviews reported a *d* of .39 (corrected for interrater reliability) for White versus Black performance for field studies (Kraiger & Ford, 1985). The *d* statistic or standardized ethnic group difference is defined as the difference in the White mean minus the Black mean divided by the sample-weighted average standard deviation of the two groups. For example, a *d* of .33 means that Whites, on average, perform or are rated approximately one third of an averaged standard deviation greater than Blacks.

Kraiger and Ford (1985) noted that subjective ratings are a function of actual performance, but such ratings may also contain biases in observation and recall of the performance. Kraiger and Ford further noted that one such set of biases could include stereotypes of Blacks held by Whites that could increase standardized ethnic group differences above true score differences in some cases. One research implication of this set of beliefs is that Black–White differences on subjective performance ratings might be larger than Black–White differences on objective performance measures.

There are other potential pressures that might influence ratings of job performance in the opposite manner. Researchers have noted that ethnicity is a highly salient and rational consideration in the evaluation of performance in organizations (Kraiger & Ford, 1985; Mobley, 1982). For example, the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978) noted that performance ratings should be scrutinized for possible race and gender effects because these ratings may serve as criteria in validation studies. Further, these devices might be used as selection devices in their own right when ratings are used in promotion decisions (Mobley, 1982). There may also be pressures to maintain and promote a diverse organization. The result of these pressures might be to motivate the rater to either intentionally or unintentionally minimize the influence of ethnicity (Mobley, 1982). This set of pressures may, or may not be, powerful enough to offset any biases inherent in subjective ratings (assuming their presence inferred in previous work). One implication of these pressures to minimize group differences is that average levels of standardized ethnic group differences for objective measures of performance might be similar to standardized differences for subjective measures of performance. There are also more complex models of rater beliefs that we discuss later under Research Needs. Unfortunately, we were not able to address a number of such issues (including a rater–ratee interactions) with our database.

In addition to the Kraiger and Ford (1985) meta-analysis, another large study of a civilian database examined Black–White differences in performance ratings (Waldman & Avolio, 1991). In a study of over 20,000 individuals selected using the General Aptitude Test Battery (GATB), Waldman and Avolio (1991) computed the Black–White performance *d* to be .35. Perhaps of most interest, when the researchers controlled for general mental ability, education, and job experience using multiple regression, the variance in performance associated with ethnicity dropped from 3.0% to 0.3%. This suggests that several job-relevant factors may at least partially explain Black–White differences in job performance. Waldman and Avolio cautioned readers that a portion of the GATB was used as a measure of general mental ability, and to the extent that this instrument is culturally biased, some bias may have been removed from the Black–White performance differences. It is also interesting to note that the rater–ratee interaction "added negligible variance" to predicting performance when race of the ratee was entered first (p. 899). Given an increased amount of data available to readdress the issue of ethnic differences on performance ratings, we pose our first research question: What are the standardized ethnic group differences for ratings of job performance for Blacks versus Whites and for Hispanics versus Whites?

Studies Examining Multiple Types of Criteria

There are several studies that examine multiple types of criteria for ethnic group differences in performance. One important theme within this stream of research is that comparing objective and subjective indicators of performance yields important insights into both the size of ethnic group differences on various indicators, and it also yields important insights into the potential bias in subjective indicators of performance. Although there may be some bias in objective measures (J. K. Ford et al., 1986), the level of bias is generally thought to be less than that in subjective measures (J. K. Ford et al., 1986; Rotundo & Sackett, 1999; Schmidt & Hunter, 1981). In fact, some researchers have interpreted objective differences as evidence that supports the contention that job performance differences between Blacks and Whites are "real" (Schmidt & Hunter, 1981, p. 1131).

In a seminal study in this area, three researchers investigated the relationship between ethnicity and several types of criterion measures for Black and White ratees, and compared objective and subjective ratings within each category (J. K. Ford et al., 1986). J. K. Ford et al. (1986) first conceptually developed five different job performance areas. Due to a small number of studies available at that time, they were forced to combine various performance measures into three categories of criteria and examined Black–White differences for both objective and subjective measures within each category. The categories were as follows: performance indicators (e.g., units produced, shortages, accidents, customer complaints), absenteeism (e.g., absenteeism, lateness), and cognitive indicators (e.g., training success and job knowledge tests).

J. K. Ford et al. (1986) found that the ethnicity–performance relationships for subjective measures of performance indicators and absenteeism were larger than the objective measures in the same category. After we converted the reported point-biserial correlations to standardized ethnic group differences, the observed *d*s for performance indicators were .25 (N = 4,287) and .44 (N = 4,130) for objective and subjective measures. For absenteeism, the *d*s were .17 (N = 21,510) and .23 (N = 2,221). In contrast, the objective *d* for cognitive indicators was larger than the subjective *d*—.63 (N = 3,389) and .41 (N = 2,782), respectively.

We believe there is an opportunity to gain further knowledge about ethnic group differences in job performance in several ways. First, one could focus on more homogeneous categories of job performance. After being forced to combine performance categories, J. K. Ford et al. (1986) were left with somewhat heterogeneous performance categories. For example, performance indicators included units produced, shortages, accidents, customer complaints, and work sample tests; absenteeism included both absences and lateness. Second, one could also focus on clearly defining the ethnic groups of interest by conducting analyses on Blacks versus Whites and Hispanics versus Whites separately. There was one study in previous analyses that included a sample labeled as "Black" that included both Blacks and Hispanics (e.g., Feild, Bayley, & Bayley, 1977). Finally, one could gather more data. We note later in this article how we were also able to add 26 new articles and technical reports to the 16 articles meta-analyzed by J. K. Ford et al.

There has also been a subsequent large-scale study of the performance of ethnic groups in the military (Pulakos, White, Oppler, & Borman, 1989). This study is notable for a number of reasons. First, the raters were provided with extensive training on the purpose of ratings and on how to prevent extraneous material from influencing ratings. Second, the authors noted that "racial bias may be less prevalent in the military environment" (Pulakos et al., 1989, p. 779). Third, the authors did not correct for unreliability in the criterion because employment decisions (e.g., promotions) were made with the "unreliable" data. This study was one of the first major studies to examine data for both Blacks and Hispanics.

Overall, ratings of task proficiency and job effort (a composite of multiple measures) in Pulakos et al. (1989) indicated that Whites scored higher than Blacks ($d = .23$) but there was little difference for the composite of personal discipline, and Blacks scored higher than Whites on the dimension of military bearing. The value of *d* for task proficiency and job effort was somewhat smaller for task performance relative to Kraiger and Ford's (1985) corrected *d* of .39 (we derived an uncorrected *d* of .33). This may be a function of using a common systematic selection system that has a screen for cognitive ability (i.e., the Armed Forces Vocational Aptitude Battery) or lower levels of ethnic biases in performance ratings. Pulakos et al. also cautioned that results of this military study might not generalize to civilian organizations. Hispanic scores on task proficiency and job effort were virtually identical to Whites ($d = -.01$). We are unaware of any other large-scale or meta-analytic studies on Hispanic–White differences in job performance.

A less well-known meta-analysis also examined different measures of job performance. Bernardin (1984) meta-analyzed the literature and found that Black–White differences were moderated by the type of performance measure. He used an intuitively appealing typology of performance measures of job knowledge measures, objective measures, performance ratings, and work samples. He found *d*s of .42 for job knowledge ($K = 7$), .05 for objective measures ($K = 14$), .23 for ratings ($K = 29$), and .48 for work samples ($K = 8$). Sample sizes (*N*) were not available. Corrected by estimates of reliability (using the values of .6 for job ratings and objective data and of .8 for job knowledge tests and work sample tests), the *d*s were .47, .06, .30, and .54, respectively.

Bernardin's taxonomy of performance measures highlights two issues. First, one may wish to consider work sample tests as a separate category. This may be justified because of an emphasis on maximum versus typical job performance (Dubois, Sackett, Zedeck, & Fogli, 1993). Maximum performance generally occurs when there is a work sample test administered to applicants or incumbents when they know they are being observed. For example, Dubois et al. (1993) hypothesized that Black–White performance differences would be lower on typical performance than on maximum performance. However, they found that the standardized ethnic group differences were larger for typical performance (*d* of .63 for speed of performance and *d* of .45 for accuracy of performance) than for maximum performance (*d*s of .12 for both dimensions of performance). Further, work sample tests typically involve fairly standardized material or situations presented to incumbents as well as standardized procedures that are used to score work sample measures (e.g., Pulakos & Schmitt, 1995). Such standardization of stimuli and scoring is probably more standardization than what is found in typical performance ratings. Thus, work sample tests eliminate variance because of differential opportunities to observe candidates and recall their performance. On the basis of these factors, we believe any system of performance criteria should clearly delineate work sample tests as a separate category.

The work of Bernardin (1984) and J. K. Ford et al. (1986) also showed that job knowledge tests may have different ethnic group differences than other categories of job performance (e.g., ratings or objective measures indicators). Larger *d*s may partially be a function of the relationship between performance of job knowledge tests and cognitive ability (Hunter, 1983a; Ree, Carretta, & Teachout, 1995) given that there are ethnic differences in cognitive ability test scores (e.g., Roth, Be Vier, Bobko, Switzer, & Tyler, 2001; Sackett & Wilk, 1994). All of this information leads us to ask our second research question: What are the standardized ethnic group differences for "other measures" of job performance, such as job knowledge tests, work sample tests, and absenteeism?

In addition, we return to the work of J. K. Ford et al. (1986) to revisit the issue of objective versus subjective performance measures. Our third research question is: How do objective and subjective measures of job performance compare in terms of the magnitude of Black–White and Hispanic–White differences?

Defining the Performance Domain

A key feature of any meta-analysis examining ethnic group performance differences is how to define and conceptualize the performance domain (J. P. Campbell, 1990). Work by Cascio (1997) provided a very useful typology of criterion measures. Cascio's typology includes measures of output/quantity, quality, lost time, turnover, training success (tests or ratings of proficiency), promotability, ratings of performance, and work samples. We added measures of safety and organizational citizenship behaviors to this list to make sure we searched and analyzed as wide a range of performance indicators as we could find.

The typology developed by Cascio (1997) also provides important advantages. The list covers a large portion of the job performance domain because it includes variables such as absenteeism (lost time), turnover, and promotability, which allowed us to add some value to the field relative to previous work (e.g., Bernardin, 1984). Likewise, Cascio's typology also provides important distinctions within certain areas of the job performance domain. For example, Cascio's typology notes distinctions between measures of quantity, quality, overall ratings of performance, and work samples. These distinctions, and an increased number of studies, also allowed us the opportunity to add knowledge to the field relative to previous work (e.g., J. K. Ford et al., 1986).

Differentiation across various portions of the job performance domain was particularly relevant because we compared objective and subjective measures of performance as a result of the use of relatively homogenous types of criteria, which allowed us to minimize the amount of extraneous variance within any given category. Thus, we had an increased ability to compare objective and subjective measures of performance without "between measure" variance (e.g., without relatively objective work sample measures vs. subjective quality of work measures obscuring such differences).

The criterion typology that we used appears in Table 1. In considering the list in Table 1, note that we only allowed performance ratings in our first category entitled "performance ratings." We did not allow ranking measures or any sort of forced-choice paired comparison measures (e.g., some coefficients from Arnold, 1968; Baehr, 1976). Inclusion of such ranking-related measures could cause supervisors to be required to distribute their subordinates' performance in ways that artificially maximize ethnic group differences, even if Black and White performance means were relatively close together.

We also explored two potential moderator variables: job complexity and level of analysis. Regarding complexity, it appears important for both the validity and standardized ethnic group differences for certain variables. It appears that validity of cognitive ability tests increases as job complexity increases (Hunter & Hunter, 1984). However, it also appears that ethnic group differences for both cognitive ability tests and interviews decrease as job complexity increases (Roth et al., 2001; Huffcutt & Roth, 1998). Thus, we did not make a prediction on the exact nature of the moderating effect in this case. We also noted the methodological cautions of confounding within-job analyses with across-jobs analysis in meta-analysis (Ostroff & Harrison, 1999). Further, the results of a previous meta-analysis showed that there were smaller standardized ethnic group differences for cognitive ability for studies using a within-job versus an across-job experimental design (Roth et al., 2001). Accordingly, we ask our fourth research question: Does job complexity or within-job versus across-job analysis moderate standardized ethnic group differences?

Contributions

This article provides several contributions to the literature. First, we update the work of J. K. Ford et al. (1986) as called for by those authors over 15 years ago. In doing so, we were able to more than double the number of studies contributing coefficients to our analysis. More detail on this is presented below in the Method section. Second, we based our analyses on more homogenous types of job performance. Within these types of job performance, we maintain the distinctions of work sample tests and job knowledge tests, and we tried to minimize the amount of extraneous variance in any one category. Third, we focused only on samples of Blacks versus Whites in our first analyses, and we then separately cumulated available data for Hispanic–White performance differences. Fourth, we explored additional moderating variables.

Method

Literature Search

We searched for articles and papers on ethnic group differences on job performance from several sources. First, we retrieved the vast majority of studies from the reference lists of previous analyses of ethnic group differences on measures of job performance (see below for more details in this issue). These analyses included the work by J. K. Ford et al. (1986), Bernardin (1984), and Martocchio and Whitener (1992). We also examined the reference list from a key meta-analysis of predictors of job performance (Schmitt, Clause, & Pulakos, 1996) because these authors reported data and references for work sample and job knowledge tests. Second, we searched a number of databases including PsycINFO from the American Psychological Association, Abstracted Business Information (known as ABI Inform), Dissertations Abstracts, and Educational Resources Information Center (known as ERIC). We also wrote to several individuals working in the area or to individuals who we believed might have relevant data.

Criteria for Inclusion

We developed a list of seven criteria for inclusion. Our intent in constructing this list was to minimize the extraneous variance in our analyses due to heterogeneous groups of subjects and dependent samples. First, studies could not combine different ethnic groups into a single,

Table 1
Types of Job Performance Measures

- Direct observation or measurement of job performance
  - Overall ratings of job performance
  - Measures of quality: ratings of quality, objective measures of work product, errors, complaints
  - Measures of quantity: output, units produced, ratings of quantity, volume of sales
- Other measures of job performance
  - Job knowledge measures: ratings of job knowledge or tests of job knowledge used to assess mastery of training material
  - Work sample tests: tests designed to directly simulate the tasks of the job, the environment of the job, or both
  - Lost time: objective measures or ratings of absenteeism and tardiness
  - Turnover: objective measures or ratings of length of time on the job and termination
  - Training success: on-the-job ratings of training performance that include more of the "criterion space" than just measures of job knowledge
  - Promotion: objective and subjective measures of promotions, ratings of promotion potential, salary, and salary increases
  - Safety: number or ratings of accidents, number of safety violations
  - Organizational citizenship behaviors: ratings or objective measures of workers going beyond task requirements of the job description

heterogeneous category. In some studies we found that researchers combined certain groups such as Hispanics and Blacks into a single minority category (e.g., Feild, Bayley, & Bayley, 1977; J. K. Ford & Kraiger, 1993; Guinn, Tupes, & Alley, 1970; Hattrup & Schmitt, 1990; Kesselman & Lopez, 1979; Lance, Johnson, Southitt, Bennett, & Harville, 2000; Morstain, 1984; Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977; Toole, Gavin, Murdy, & Sells, 1972). We also did not consider coefficients from studies in which there was a heterogeneous mix of groups entitled "Nonminority" or non-Black (e.g., some analyses in Lance et al., 2000).

Second, we required that ratings of performance were made by others and not the research subjects themselves. This required an individual outside of the subject to report performance information to minimize variance in performance attributable to impression management and self-defense mechanisms. For example, we required ratings of job performance be made by either supervisors or peers. This excluded a few coefficients based on self-rated job performance (e.g., Grant-Vallone, 1998; Yu, 1998), but there were not enough of these studies to perform meaningful moderator analysis.

Third, studies must have reported data in which there were both Blacks and Whites (or Hispanics and Whites) in the same organization. Studies that compared Black performance in predominantly Black organizations to White performance in predominantly White organizations were excluded from analysis (e.g., Study 4 in Kirkpatrick, Ewen, Barrett, & Katzell, 1968).

Fourth, results of studies could not have been subject to range enhancement (Hunter & Schmidt, 1990), also called reverse range restriction (see Bobko, 2001). Range enhancement might occur if only the top and bottom thirds of the performance distribution were sampled, and *d*s would be artificially inflated. This state of affairs does not practically reflect typical personnel selection situations. Precluding studies with range enhancement meant that Baehr et al. (1971) could not be used in analyses. This is a judgment call in meta-analysis, and we note that this study was used by J. K. Ford et al. (1986).

Fifth, data must be computed from independent samples. We chose not to include the data from two studies of the U.S. Employment Service database analyzing the GATB (Rotundo & Sackett, 1999; Sackett & DuBois, 1991) because we had already included a similar database in which the vast majority of the same subjects were analyzed (Waldman & Avolio, 1991). Further, we were particularly careful when examining articles with objective and subjective measures computed from the same sample. In this case, the objective measure was entered into the overall analysis for that measure (assuming objective measures had lower levels of bias), but both measures were retained for use in our moderator analysis of objective versus subjective measures of a given type of performance.

Sixth, we eliminated studies with explicit mention of affirmative action programs in the text of articles or technical reports (we further required that we be able to see the presence of such programs in subsequently reported statistics before eliminating these studies). We were concerned that affirmative action programs might result in different hiring standards for ethnic groups and that this practice might artificially increase ethnic group differences. For example, we noted that Kraiger (1981) reported the presence of an affirmative action program in which an organization hired "only the highest scoring White applicants ('the cream of the crop') while hiring nearly all Black applicants to meet affirmative action goals" (p. 49). The author furthered characterized this as applying hiring standards to one group and that "another group was hired indiscriminately" (p. 49). Such hiring standards were associated with a supervisory performance rating *d* of .95 (approximately triple the magnitude of our later findings) and a job knowledge *d* of 1.60 (more than double the magnitude of our later findings). Such ethnic performance differences are not likely to be typical of most organizations and are not included in our analysis (Bartlett et al., 1977; Kraiger, 1981).

Seventh, we eliminated output (i.e., number of arrests) and quality (i.e., disciplinary actions) coefficients from one police study because of a potential confound in computing the value of *d*. Specifically, these measures of performance were noted in the study to be problematic because Black individuals were assigned to primarily Black neighborhoods (with lower socio-economic status) and White individuals to White neighborhoods (Mills, 1990). That is, individuals were clearly given differential opportunity for success. The number of arrests and disciplinary actions were influenced by what some researchers would label *opportunity bias* (Goldstein, 1974). For example, the number of complaints filed against Black individuals were extremely high and likely to be influenced by where they were assigned rather than by their own behavior.

Calculation of ds

We calculated *d* from means and standard deviations whenever possible. If means and within-group standard deviations were not available, we derived *d* from *F* or *t* values. If data were reported in multifactor tables, we were careful to find the *d* associated only with the main effect for ethnicity. For example, the data might be reported in a 2 × 2 table (e.g., gender by ethnicity as per Mills, 1990). In such cases, we were careful to estimate the variance due to gender and recalculate the appropriate standard deviations for ethnicity. In this way, we did not overestimate the values of *d* by having variance from another source partialled out of within-group standard deviations.

We also formed composites when two measures of somewhat different portions of a given performance domain were available (and the intercorrelations were available). For example, cash shortages and overages for a sample of bank tellers (Bass & Turner, 1973) were combined into a unit-weighted composite representing objective quality of work. This precluded the inclusion of dependent samples and gave us the most accurate picture of ethnic group differences within a given performance area.

Last, we only calculated *d* if there were 10 or more minorities in the sample. This resulted in a loss of data from several samples (e.g., Cascio & Phillips, 1979; Neidt, 1968), but we preferred only including standardized differences that would be associated with reasonably stable effect sizes.

Overlap With Previous Meta-Analyses on Performance Differences

Two other sets of researchers conducted meta-analyses that examined ethnic group differences on a variety of measures of job performance. We examined the list of studies incorporated in the analysis conducted by J. K. Ford et al. (1986). They cited 16 studies that yielded coefficients. We were able to obtain 15 of these studies, although 5 did not meet our criteria for inclusion. We added 26 additional studies to our database that were not in J. K. Ford et al.'s database. It was more difficult to ascertain the overlap with the work of Bernardin (1984) because a list of studies that yielded data for the meta-analysis was not specified (as such conventions were not as commonplace in 1984 as they are today). We compared the reference list of all cited works in Bernadin's article with our list of studies that yielded data for our meta-analysis. We were able to note 28 studies in our database that were are not included in his analysis. Thus, it appears there is a substantial amount of data to add to previous analyses.

Coding the Data

To maximize the quality and accuracy of our dataset, the first two authors independently coded all study characteristics. First, they coded studies by the category of performance measure(s) used (see Table 1). Second, they coded for the level of job complexity by using the scale developed by Hunter (1983b), which ranged from low (e.g., mail sorter), to low–medium (e.g., truck driver), to medium (e.g., skilled crafts), to medium–high (e.g., computer trouble-shooter), to high (e.g., executives, scientists). Third, they coded whether the study was done on one job (within job) or multiple jobs or multiple organizations (across job). Fourth, they coded the standardized ethnic group difference (*d*) and sample size.

There was substantial coding agreement for both categorical and continuous variables. The coders initially agreed 99% of the time on the category of performance measure (e.g., performance rating, job knowledge test). They initially agreed 86% of the time on complexity categories (and never disagreed by more than one category), and they agreed on 86% of the initial coding for within-job versus across-jobs judgments. The correlation for initial calculations for sample size was .99 and the correlation for the size of *d* was .96. When disagreement occurred, they reached agreement on final codes by consensus.

The authors also noted if the study involved a police organization. Validities for police organizations are often lower than for other organizations due to the relatively infrequent opportunities that supervisors have to observe their subordinates (e.g., McDaniel, Whetzel, Schmidt, & Maurer, 1994). If supervisors do not have an opportunity to observe their subordinates, standardized ethnic group differences for job performance might also be problematic. Thus, this coding allowed us to control the variance across studies associated with police samples.

Correcting d

We computed both observed *d*s and corrected *d*s. We computed the observed *d* for insight into the influence that performance differences might have for adverse impact on operational business decisions such as promotions and salary increases. We report the *d* corrected for measurement reliability for insight into actual differences in performance that are not influenced by measurement error.

Determining the reliability values for corrections was somewhat difficult. We found that other researchers had corrected for unreliability by using a reliability of .70 for performance ratings, .80 for cognitively related measures, and .60 for absenteeism and objective measures (e.g., J. K. Ford et al., 1986).

We also examined our own data for reliability values and found only four studies that provided interrater reliability for performance rating measures. These values averaged .595. This converged with other meta-analytic corrections using a value of .60 for performance ratings (see examples in Bobko, Roth, & Potosky, 1999), thus we used the value of .6. We note that the value was somewhat more conservative than the interrater reliability value of .52 suggested by Viswesvaran, Ones, and Schmidt (1996).

We found no guidance in our data for setting the reliability values for measures of absenteeism and objective measures of job performance. We chose to set the corrections for all objective measures of job knowledge, work samples, absenteeism, and objective measures of quality and quantity at .80.

We corrected our estimates for reliability in two different ways. First, we corrected results with only objective or subjective measures using the reliability values stated above. For example, the mean observed *d* for subjective measures of quality is .20 (see Table 3). We corrected the value of .20 for measurement error with a reliability of .6 to yield .26. Second, we individually corrected coefficients when an analysis contained both objective and subjective measures. For example, we individually corrected each of the 8 objective quality coefficients (summarized in Table 3) by using the value of .8 and then individually corrected each of the 10 subjective measures of quality (by using the value of .6). We then averaged the 18 values to arrive at the corrected value of .25 reported in Table 2.

Results

We first report Black–White results and later report Hispanic–White analyses for which we had sufficient data. We generally did not report results for any type of criterion unless there were four or more coefficients available for analysis within a category noted in Table 1. We did make an exception to this guideline because we did report a meta-analytic estimate for training measures. This was done to allow readers to examine this category and its possible impact if it was combined with various measures of job knowledge to provide some overall estimate of training–job knowledge differences. We also did report some moderator analyses within a category with four coefficients (from Table 1). For example, we report analyses of objective versus subjective measures of job performance with only two coefficients within each category in Table 5 to allow readers to see if the pattern of results supports or disconfirms the third research question. We urge caution in the interpretation of some of these small *K* estimates if one is looking for highly stable point estimates of a particular phenomenon. Finally, we conducted a number of analyses removing coefficients from police studies, but this did not change our results in any meaningful way. Thus, we report all results with police studies included.

Overall Analyses of Performance Differences

Our first research question addressed standardized ethnic performance differences for ratings of job performance, and our

Table 2
Black–White Differences in Job Performance

Measure	*d*	*K*	N_{Total}	N_{White}	N_{Black}	90% CI	$d_{\text{corrected}}$	PVA (%)
Ratings								
Overall ratings	.27	37	84,295	62,073	22,222	25, .29	.35	35
Overall with no military	.31	36	46,183	35,023	11,160	.28, .33	.40	43
Overall no large *N*s	.26	33	9,141	6,693	2,448	.23, .30	.34	76
Quality & quantity measures								
Quality ratings	.21	15	3,613	2,387	1,226	.14, .27	.25	89
Quantity ratings	.21	8	1,268	925	343	.03, .40	.26	69
Other measures								
Job knowledge	.48	12	2,460	1,577	883	.36, .58	54	30
Work samples	.52	10	3,651	2,260	1,391	.39, .66	.59	18
Absenteeism	.19	11	2,376	1630	746	.11, .26	.22	100
On-the-job training	.14	2	132	75	56	−.03, .31	.18	100
Promotion	.31	7	1,404	1,168	236	.20, .42	.38	100

Note. PVA = the percentage of variance accounted for by sampling error.

Table 3
Comparison of Objective and Subjective Measures of Job Performance

Measure	d	K	N_{Total}	N_{White}	N_{Black}	90% CI	$d_{\text{corrected}}$	PVA (%)
Quality measures								
Objective	.24	8	2,538	1,632	906	.17, .30	.27	100
Subjective	.20	10	1,811	1,262	549	.12, .28	.26	100
Subjective, no part time	.14	9	1,580	1,063	517	.10, .20	.18	100
Quantity measures								
Objective	.32	3	774	613	161	−.06, .72	.35	84
Subjective	.09	5	494	312	182	−.06, .24	.12	100
Job knowledge								
Objective	.55	10	2,027	1,315	712	.42, .68	.61	34
Subjective	.15	4	1,231	793	438	.08, .23	.19	100
Absenteeism								
Objective	.23	8	1,413	1,005	408	.12, .32	.26	90
Subjective	.13	4	642	377	275	.09, .17	.17	100

Note. Objective measures of performance were corrected for attenuation using the value of .8, whereas subjective measures were corrected by using the value of .6. PVA = the percentage of variance accounted for by sampling error.

second question addressed differences for other measures of job performance. Table 2 presents results to address these questions. The first row in Table 2 represents ratings of overall job performance. This analysis contains a larger number of total subjects than previous analyses. Our overall results suggest that Whites and Blacks differ by an observed *d* of .27, and a 90% confidence interval around this value varies from .25 to .29. The value of *d* increases for organizations that are not military in their nature (to an observed *d* of .31). Eliminating all of the samples with an *N* of 2,000 or more does not appreciably change the estimate of the overall *d* (.26), showing that large samples did not have a disproportionate influence on results. Considering all of the results, it appears that Whites are rated higher than Blacks by somewhat less than one third of an averaged standard deviation for the uncorrected value of *d*. Our observed values are quite close to some previous analyses. They are somewhat smaller than those found by Kraiger and Ford (1985; $d = .33$) and slightly larger than those found by Bernardin (1984; $d = .24$).

Results for measures of quantity and quality are also reported in Table 2. Overall analyses for measures of both quantity and quality show observed standardized differences of .21. Comparisons of objective and subjective measures are presented later.

Job knowledge and work sample test results are of special interest for three reasons. First, these two types of “criteria” can also be viewed as predictors. That is, applicants could be selected on the basis of job knowledge or work sample tests (e.g., Salgado, Viswesvaran, & Ones, 2001). Second, work sample tests are often standardized in both the presentation of material to individuals and their scoring of that material for individuals. Thus, as noted earlier, they eliminate certain sources of variance in criterion measurement such as differential exposure to ratee behavior. They also can focus more on maximum versus typical performance. Third, we suggested earlier that job knowledge tests were probably at least moderately conceptually related to cognitive ability.

Table 2 shows that the overall observed *d* for all measures of job knowledge is .48. This value is somewhat higher than the values for ratings of overall job performance. This may be a function of the cognitive demands of such tests. It is also important to note that the above meta-analytic figures include both objective and subjective measures of job knowledge. Readers interested in results for either type of measure are referred to Table 3.

Work sample tests were associated with an observed *d* of .52 corrected to .59. This *d* was slightly larger in size than for job knowledge tests, although we note that confidence intervals overlap. It is unclear how this *d* compares with results in J. K. Ford et al. (1986) because these researchers were forced to include work sample tests with other measures of job performance due to the limited sample size of studies at that time (personal communica-

Table 4
Moderator Analyses of Overall Job Performance Ratings

Moderator	d	K	N_{Total}	N_{White}	N_{Black}	90% CI	$d_{\text{corrected}}$	PVA (%)
Job complexity								
Low complex	.32	5	994	600	394	.16, .48	.41	50
Low–medium complex	.27	20	11,916	8,973	2,943	.23, .31	.35	89
Medium complex	.32	6	11,375	10,234	1,141	.25, .39	.41	56
Medium complex; no large *N*	.31	5	942	571	371	.18, .45	.40	63
Within vs. across jobs								
Within jobs	.23	19	3,855	2,601	1,254	.18, .29	.30	100
Across jobs	.27	16	79,943	59,050	20,893	.24, .30	.35	22
Across jobs; no military	.32	15	41,831	32,000	9,831	.28, .36	.41	25

Note. Objective measures of job knowledge and work samples were corrected of attenuation by using the value of .8, and subjective measures of job knowledge were corrected by using a value of .6. PVA = the percentage of variance accounted for by sampling error.

Table 5
Hispanic–White Differences on Measures of Job Performance

Measure	d	K	N_{Total}	N_{White}	N_{Hispanic}	90% CI	$d_{\text{corrected}}$	PVA (%)
Overall ratings								
Overall	.04	11	46,530	43,909	2,621	.00, .08	.05	67
Job complexity								
Low–medium	.07	6	7,499	6,750	749	.02, .11	.09	100
Medium complex	.16	3	10,213	9,992	221	−.01, .33	.21	50
Other measures								
Job knowledge	.47	3	977	705	272	.12, .86	.53	11
Objective	.67	2	698	500	198	.36, 1.02	.75	20
Subjective	.04	2	621	446	175	.00, .08	.05	100
Work samples	.45	4	1,197	866	331	.24, .68	.50	23

Note. PVA = the percentage of variance accounted for by sampling error.

tion, J. K. Ford, 22 May 2002). Our *d* of .52 is quite similar to the observed *d* of .48 found by Bernardin (1984). It is interesting to note that sampling error only accounted for approximately 18% of the variance in this category. This is a relatively small portion of the variance relative to other categories and may partially be a function of various work samples measuring a variety of constructs.

It is also interesting to compare both of these *d*s with another meta-analysis. Schmitt et al. (1996) meta-analyzed the standardized ethnic group differences for a single, combined category that included job knowledge, work sample, and situational judgment tests. Types of tests were combined due to limited sample sizes. Schmitt et al. reported a *d* of .38. Our results allow researchers to have specific point estimates for job knowledge tests and work sample tests as criterion measures.

Table 2 also shows our results for measures of absenteeism, training success, and promotion. The overall observed *d* for absenteeism was .19. We note two things about this analysis: All studies are of measures about absenteeism. There were only two studies that we found of lateness, and we did not include them in this analysis in order to minimize extraneous variance. The *d* for studies using on-the-job measures or a combination of on-the-job and paper and pencil measures to assess training success was .14, although we found only two studies in this category. Table 2 also shows our analysis for promotion related measures. These included measures of promotability, salary increases, and actual promotions. Thus, it is a very heterogeneous category and we urge caution in its interpretation. The overall *d* was .31.

Unfortunately, there were not enough data for turnover, organizational citizenship behavior, or safety to analyze. We were disappointed that we could find only one study of organizational citizenship behavior (similar to Borman, Penner, Allen, & Motowidlo, 2001; Podsakoff, MacKenzie, Paine, & Bachrach, 2000). A unit-weighted composite of altruism and conscientiousness extra-role behaviors in one study showed a *d* of −.10, indicating that Blacks scored higher (Grant-Vallone, 1998). The same study reported a *d* of −.07 for Hispanic–White standardized differences.

Analysis of Objective Versus Subjective Measures of Performance

Our third research question addressed the relationship between objective and subjective measures of performance.

Quantity and quality measures. Table 3 first presents the results for measures of quality and measures of quantity. For quality, the objective measures observed *d* is .24 and the subjective measures observed *d* is .20 and confidence intervals do overlap, as one might expect from the magnitudes of the two point estimates. When corrected for unreliability, the *d*s become .27 and .26, respectively. We also report the results for subjective ratings of quality without the influence of one study of part-time workers to allow readers to gauge its influence on overall calculations. The *d* for this study was notably larger than the other studies (and was identified as an outlier–influential coefficient by the sample adjusted meta-analytic deviancy [SAMD] statistic). The values of objective and subjective measures are quite similar and are not consistent with the assumption that subjective measures of performance are associated with larger *d*s than are objective measures.

The pattern for comparing objective and subjective measures of quantity is clearer but less definitive. The objective *d* is .32, although it is based on only three studies, and there is a wide confidence interval around it indicating substantial variance in the *d*s of the three studies. In comparison, the subjective *d* is .09, although it is based on only five studies. Given the available data, we would not suggest that the objective *d* is definitely larger; however, there is certainly no evidence to suggest the objective *d* is smaller. In sum, analysis of indicators of quality and quantity suggest that objective *d*s are not less than subjective *d*s.

Job knowledge tests. Table 3 also shows that objective measures of job knowledge are associated with higher *d*s than subjective measures. The *d*s are .55 and .15 uncorrected for measurement reliability (.61 and .19 corrected for measurement reliability). We note that even though there were only 12 independent job knowledge samples, there were two studies that reported both objective and subjective measures of job knowledge (see J. T. Campbell, Crooks, Mahoney, & Rock, 1973). This allowed us to have a total of 14 coefficients in our moderator analysis. One study that contained both objective and subjective measures of job knowledge for the job of medical technicians ($N = 456$) was associated with a *d* of .44 for an objective test and of .06 for a subjective rating. A second study of cartographic technicians ($N = 342$) resulted in *d*s of .42 and .23, respectively. Overall, these particular results are quite consistent with those of J. K. Ford et al. (1986) with their general category of cognitive indicators.

Absenteeism. The objective *d* for absenteeism was .23, and the subjective *d* was .13. Interestingly, this objective versus subjective analysis fits a pattern in which objective measures appear to be at least as big as, and sometimes larger than, subjective measures.

Moderators

We also examined the moderators of job complexity and within-job versus across-job samples for ratings of job performance in Table 4. We are unaware of any other meta-analysis in this area that has reported such analyses. Our analyses show somewhat mixed results in terms of a moderating effect. The mean *d* for low complexity jobs was .32, but it was based on only five studies and there was a large amount of nonsampling error variance in study results. Low–medium complexity jobs are associated with an observed *d* of .27, and a category combining medium complexity and medium–high complexity jobs was associated with an average *d* of .32. There were no high complexity jobs in our database. Perhaps the clearest conclusion is that there are positive standardized ethnic group differences at all levels of complexity analyzed in Table 4.

The results of our within versus across-job studies suggest a possible moderating effect. Within-study jobs are associated with a *d* of .23, whereas across-job studies without military samples are associated with a *d* of .32, although confidence intervals overlap. We conducted analyses with and without the large military sample size military study by Pulakos et al. (1989) because the authors noted that the results of this study may not generalize to other types of organizations as noted above.

Hispanic Results

There were many fewer studies of Hispanic–White performance comparisons than for Black–White comparisons. Table 5 shows the results for these analyses. Overall, the standardized ethnic group differences for performance ratings of Hispanics is .04, smaller than that for Blacks. Table 5 also shows an analysis of job knowledge and work sample studies. In both cases, the results are quite tentative because sample sizes and numbers of studies are small. The overall observed *d* for job knowledge measures is .47; the objective *d* is .67, whereas the subjective *d* is .04. The last two coefficients are based on only two studies each. There was one study of cartographic technicians ($N = 342$) that provided both an objective and subjective measure of job knowledge and the *d*s were .40 and .03, respectively (J. T. Campbell et al., 1973). Work sample studies were associated with a *d* of .45 on the basis of four coefficients. The analyses of work samples and job knowledge tests also differs in comparison with the estimates from the estimates of previous work (Schmitt et al., 1996), although again we caution the reader about our sample sizes. Those researchers found a *d* of .00 for Hispanic–White comparisons in a heterogeneous category of job knowledge tests, work sample tests, and situational judgment tests. The differences in results may be attributed to smaller, more homogenous categories (e.g., only work samples or job knowledge measures), to our criteria for including studies in this analysis, or to the fact that Schmitt et al. (1996) focused on three major journals in applied psychology (because their purpose was to compare across different types of predictors of performance), whereas we covered a wider variety of journals, technical reports, and dissertations.

In terms of job complexity, results show that low–medium complexity jobs are associated with a *d* of .07, whereas medium and medium–high complexity jobs as a single category are associated with a *d* of .16, although the estimate is based on only three studies.

Discussion

Main Effects

The results of this study reinforce some beliefs and change others. For Black–White comparisons, the overall results show a standardized ethnic group difference for job performance ratings of approximately one third of a standard deviation (when corrected for criterion reliability), and this is quite similar to Kraiger and Ford (1985). We also had similar results for one of three types of performance measures used by J. K. Ford et al. (1986). Specifically, we found larger *d*s associated with objective measures of job knowledge than with subjective measures of job knowledge. It is also interesting to note that the overall analysis of objective versus subjective criteria (collapsed across criterion categories) by Ford et al. showed corrected point biserial correlations of .209 for objective criteria and of .204 for subjective criteria. We return to the issue of similarity of the objective and subjective measures below.

There were important differences between the findings of our meta-analysis and previous work. Fortunately, we had a larger sample of studies and were able to conduct analyses within more, but not completely homogeneous, categories of performance. We found a clear pattern of *d*s in which objective measures of performance were generally of similar or larger size relative to subjective measures of performance across a variety of criterion types. For example, objective measure *d*s were similar or somewhat larger in size for measures of quality, quantity, and absenteeism (although sample sizes were small in some cases). Thus, it appears that indicators aimed at measuring performance and absenteeism are part of the pattern in which objective measures showed similar or somewhat larger Black–White standardized differences.

Our pattern of results has important implications for the study of standardized ethnic group differences. Our results do not support the position that subjective measures have more potential for bias than objective measures. Instead, we found the opposite. This is important because J. K. Ford et al. (1986) noted that some researchers (not necessarily including themselves) have called for the increased use of objective measures to minimize Black–White differences based on the implicit assumption that objective measures are less prone to bias than subjective measures. Our results are more consistent with a position that there may be some pressure to minimize ethnic group differences on raters (e.g., Mobley, 1982). We also note that our results are consistent with the results of Ford et al. in that criterion measures are not substitutable; different types of criteria are likely to have different *d*s (e.g., ratings vs. job knowledge tests).

We also presented estimates of Hispanic–White differences in measures of job performance. Standardized differences were smallest for overall ratings of performance (.04), whereas differences for job knowledge tests and work sample tests were larger (.47 and .45, respectively). Within the category of job knowledge measures, there were two studies using objective tests that were associated with a larger average *d* (.67) than the two studies using subjective ratings (.04). We urge caution in interpreting the last two *d*s given limited sample size.

Moderators of Ratings of Job Performance

For Black–White comparisons of job complexity, low–medium complexity jobs were associated with slightly lower *d*s than were medium and high–medium complexity *d*s. However, the cross-cell differences were modest, making interpretation of any linear effect difficult. It was interesting to note that there were positive *d*s in all three categories of complexity, indicating that there were differences in job performance at all three of these levels of complexity. For Hispanic–White comparisons, low–medium complexity jobs were associated with smaller *d*s than were medium and high–medium complexity jobs, although confidence intervals overlapped.

There may be some moderating effect due to study design such that studies analyzing a single job for an organization are associated with smaller *d*s than across-job studies. Such results are somewhat encouraging because they suggest that within-job standardized differences for ratings may be somewhat less than previously thought. This may also be relevant information for researchers seeking to model ethnic group differences within a given organization and within a given job.

The results of our meta-analysis are also able to partially disentangle the Black–White *d*s for job knowledge versus work sample measures. The estimated uncorrected Black–White *d* for job knowledge tests is .48, is .52 for work sample tests, and the confidence intervals overlap. We should note that both of these figures were based on job incumbents and are appropriate for estimating the *d* for job performance differences or perhaps the *d* for promotional (not entry-level) tests. We also note similar Hispanic–White *d*s of .47 for job knowledge measures and of .45 for Hispanic–White standardized differences on work samples. It is also important to note that work sample tests may assess a variety of constructs (e.g., mental abilities vs. interpersonal abilities) and that the *d* for such tests could be influenced as much or more by construct as method of measurement.

Limitations

Every study, regardless of its scope, has limitations. We note three. The first was the lack of available data for some types of performance indicators in our expanded categorization scheme. It was quite difficult to find enough studies to conduct some analyses, such as analyses of Hispanic–White differences for criteria such as job knowledge and work sample tests or Black–White differences for organizational citizenship behaviors and indicators of safety. Sample sizes were also quite small for a number of moderator analyses. For example, it was difficult to compare objective versus subjective measures of performance within many relatively homogeneous types of performance. Although one could partially overcome this problem by looking across criterion types, it was a substantial limitation.

The second limitation was that we could not directly examine the amount or source of bias in any of the criteria. This was potentially most problematic for rating criteria that may suffer from more sources of bias. Like other meta-analyses before ours (e.g., Bernardin, 1984; J. K. Ford et al., 1986; Kraiger & Ford, 1985), we could not tease apart true score variance from bias in any direct manner. However, we were able to shed some light on issues of bias by comparing objective and subjective measures of performance (as per J. K. Ford et al., 1986).

The third limitation is that we could not investigate rater–ratee interactions. We briefly summarize this literature for the sake of complete coverage of the issues and then note our limitation. Kraiger and Ford (1985) found that White raters rate White subordinates higher ($d = .37$), whereas Black raters rate Black subordinates higher ($d = -.45$). Other researchers have shown how some peer and laboratory studies have inflated the value of -.45 (Sackett & Dubois, 1991). The general trend for both civilian and military samples is for both White and Black raters to rate White subordinates higher, although the *d* is typically larger for White raters (Dubois & Sackett, 1991; Pulakos et al., 1989; Rotundo & Sackett, 1999). A notable exception is Mount, Sytsma, Hazucha, and Holt (1997) because they found that White raters rated White ratees notably higher, and Black raters often rated Black ratees notably higher when ratings were used for developmental purposes. Unfortunately, we could not analyze this interaction effect given the nature of the data reported in our primary studies.

Future Research

There are at least two streams of research that could help us further understand ethnic group differences in job performance. The first stream of research focuses on gathering more data that is similar to this meta-analysis. More studies that could be coded for factors such as job complexity, objective versus subjective criteria, and within-job versus across-job analysis would be very useful. Within this stream, individuals might also look at the purpose for ratings (e.g., administrative vs. research). Although we tried to code this variable ourselves, we found that we were often not reasonably sure of the purpose in many studies. In many cases, the purpose was not mentioned at all. This may be a case in which a large and well-designed primary study would be helpful in answering such questions.

A second stream of research might be geared toward answering questions of why there are differences in ethnic groups for the variable of job performance. One line of inquiry might look at job related variables. Recall that Waldman and Avolio (1991) looked at cognitive ability, education, and job experience, and controlling these variables reduced the variance in performance attributable to ethnicity from 3.0% to 0.3%.

A second line of this research might focus on examination of the complex, multiple motivations and beliefs concerning bias, or potential bias, in subjective measures of job performance. Such research might examine the potentially complex sets of beliefs that raters bring to the task of rating subordinates. One set of beliefs that deserves further research is the presence and potential pervasiveness of beliefs that White raters have about Black subordinates. Simply put, how strong and pervasive are such beliefs? A second set of beliefs concerns the presence, effect size, and pervasiveness of pressures to minimize ethnic differences in performance ratings. Given the widely publicized legal cases of major Fortune 500 companies struggling with differences in job performance and promotion rates of Whites and Blacks, do raters feel pressures to minimize ethnic group differences? A third set of beliefs that deserves further research concerns beliefs held by Black raters. The authors are unaware of any substantial research that addresses whether or not such raters hold any beliefs about White subordinates, whether they have encountered bias in their own ratings of performance, and whether they feel any pressure to make up for perceived bias when rating other Blacks. Although some research looking at biases (e.g., pressure to minimize ethnic group differences) might be furthered through quantitative and qualitative questionnaires and surveys, other forms might require policy capturing or less intrusive methods given the sensitive nature of the issues involved.

A related line of research might focus on why Hispanic–White standardized group differences for ratings of job performance are smaller than Black–White differences. Although there are mean differences between Hispanics and Whites on cognitive abilities tests (Sackett & Wilk, 1994; Roth et al., 2001), the Hispanic–White performance difference appears to be much smaller than the Black–White difference. All told, there is a great deal of research that could follow this analysis, and we look forward to the increased understanding it will bring.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Arnold, B. C. (1968). *Comparison of Caucasian and Negro subgroups on criterion indices of overall job effectiveness* (AAT 6912449). Unpublished doctoral dissertation, Colorado State University, Fort Collins.

Baehr, M. E. (1976). *National validation of a selection test battery for male transit bus operators* (Technical Report No. UMTA-MA-06–011–77–1). Washington, DC: U.S. Department of Transportation.

Baehr, M. E., Saunders, D. R., Furcon, J. E., & Froemel, E. C. (1971, September). The prediction of performance for Black and White police patrolmen. *Professional Psychology,* 46–57.

Bartlett, C. J., Goldstein, I. L., Mosier, S., Hannan, R., Buxton, V., Simmons, V., et al. (1977). *An analysis of the validity of the PPA police examination for entry level selection in the Prince George's County 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.

Bernardin, H. J. (1984). *An analysis of Black-White differences in job performance.* Paper presented at the meeting of the Academy of Management, Boston.

Bobko, P. (2001). *Correlation and regression: Principles and applications for industrial/organizational psychology and management* (2nd ed.). London: Sage.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 62,* 561–589.

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship behavior. *International Journal of Selection and Assessment, 9,* 52–69.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

*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: ETS.

Cascio, W. F. (1997). *Applied psychology in human resource management.* Upper Saddle River, NJ: Prentice Hall.

Cascio, W. F., & Phillips, N. F. (1979). Performance testing: A rose among thorns. *Personnel Psychology, 32,* 751–766.

*Confidential Industry Technical Report.

*Confidential Industry Technical Report.

*Confidential Industry Technical Report.

*Distefano, M. K., Pryer, M. W., & Craig, S. H. (1976). Predictive validity of general ability tests with black and white psychiatric attendants. *Personnel Psychology, 29,* 197–204.

*Distefano, M. K., Pryer, M. W., & Craig, S. H. (1980). Job-relatedness of a post-training job knowledge criterion used to assess validity and test fairness. *Personnel Psychology, 33,* 785–793.

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

Equal Employment Opportunity Commission. (1978). *Uniform guidelines on employee selection procedures, 43,* 38290–38309.

*Farr, J. L., O'Leary, B. S., & Bartlett, C. J. (1971). Ethnic group membership as 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 non-minority production workers. *Personnel Psychology, 30,* 37–46.

Ford, J. K., & Kraiger, K. (1993). Police officer selection validation project: The multi-jurisdictional police officer examination. *Journal of Business and Psychology, 7,* 421–429.

Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. *Psychological Bulletin, 99,* 330–337.

*Ford, K. A. (1976). *Ethnic group differences in employment test-job performance relationships.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

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

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

*Gael, S., & Grant, D. L., & Ritchie, R. J. (1975). Employment test validation for minority and nonminority clerks with work sample criteria. *Journal of Applied Psychology, 60,* 420–426.

*Gael, S., & Grant, D. L., & Ritchie, R. J. (1975). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology, 60,* 411–419.

Goldstein, I. L. (1974). *Training: Program development and evaluation.* Monterey, CA: Brooks/Cole.

*Grant, D. L., & Bray, D. W. (1970). Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology, 54,* 7–14.

*Grant-Vallone, E. J. (1998). *Work and family conflict: The importance of supportive work environments.* Unpublished doctoral dissertation, Claremont Graduate University, Claremont, CA.

Guinn, N., Tupes, E. C., & Alley, W. E. (1970). *Cultural subgroup differences in the relationships between Air Force aptitude composites and training criteria* (Technical Report No. AFHRL-TR-70–35). Lackland Air Force Base, TX: Personnel Research Division.

Hattrup, K., & Schmitt, N. (1990). Prediction of trade apprentices' performance on job sample criteria. *Personnel Psychology, 43,* 453–466.

*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.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in interview evaluations. *Journal of Applied Psychology, 83,* 288–297.

Hunter, J. E. (1983a). A causal analysis of cognitive ability, job knowledge, and job performance, and supervisory ratings. In R. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257–266). Hillsdale, NJ: Erlbaum.

Hunter, J. E. (1983b). *Test validation for 13,000 jobs: An application of job classification and validity generalization to the General Aptitude Test Battery (GATB)* (USES Test Research Report No. 450). Washington, DC: U.S. Department of Labor.

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. (1990). *Methods of meta-analysis: Correcting for error and bias in research findings.* Newbury Park, CA: Sage.

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

*Kahn, E. (1977). *A study of the use of work sample criterion in test validation research.* Unpublished doctoral dissertation, University of Houston, University Park, TX.

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,* 91–107.

*Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. (1968). *Testing and fair employment.* New York: New York University Press.

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

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

Lance, C. E., Johnson, C. D., Southitt, S. S., Bennett, W., Jr., & Harville, D. L. (2000). Good news: Work sample administrator's global performance judgments are (about) as valid as we've suspected. *Human Performance, 13,* 253–277.

*Lopez, F. M. (1966). Current problems in test performance of job applicants: 1. *Personnel Psychology, 19,* 10–18.

Martocchio, J. J., & Whitener, E. M. (1992). Fairness in personnel selection: A meta-analysis and policy implications. *Human Relations, 45,* 489–506.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

*Mills, A. (1990). *Predicting police performance for differing gender and ethnic groups: A longitudinal study.* Unpublished doctoral dissertation, California School of Professional Psychology, Los Angeles.

*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.

Morstain, B. R. (1984). Minority–White differences on a police aptitude exam: EEO implications for police selection. *Psychological Reports, 55,* 515–525.

*Motowidlo, S. J., & Schmit, M. J. (1997). *Performance assessment and interview procedures for store manager and store associate positions* (Final Technical Report). Gainesville, FL: University of Florida, Human Resource Research Center.

Mount, M. K., Sytsma, M. R., Hazucha, J. F., & Holt, K. E. (1997). Rater–ratee race effects in developmental performance ratings of managers. *Personnel Psychology, 50,* 51–69.

*Neidt, C. O. (1968). *Report on the differential predictive validity of specified selection techniques within designated subgroups of applicants for civil service positions.* Fort Collins, CO: Colorado Civil Rights Commission.

Ostroff, C., & Harrison, D. A. (1999). Meta-analysis, level of analysis, and best estimates of population correlations: Cautions for interpreting meta-analytic results in organizational behavior. *Journal of Applied Psychology, 84,* 260–270.

*Plamondon, K. E., & Schmitt, N. (2000, April). *Validity and subgroup differences of combinations of predictors as a function of research design.* Paper presented at the 15th Annual Meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., & Bachrach, J. B. (2000). Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management, 26,* 513–564.

*Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48,* 289–308.

*Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of performance ratings: An examination of ratee race, ratee gender, and rater level effects. *Human Performance, 9,* 103–119.

*Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology, 74,* 770–780.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior job knowledge in complex training performance. *Journal of Applied Psychology, 80,* 721–730.

Roth, P. L., Be Vier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297–330.

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

Sackett, P. R., & Dubois, C. L. Z. (1991). Rater–ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology, 76,* 873–877.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929–954.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. Anderson, D. Ones, H. Sinangil, & C. Viswesvaran (Eds.) *Handbook of industrial, work, & organizational psychology* (pp. 165–199). London: Sage.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample versus paper and pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30,* 187–197.

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

*Schmit, M. J., & Motowidlo, S. J. (1995). *Development and validation of a situational interview and personality test for selecting sales associates* (Final Technical Report). Gainesville, FL: University of Florida, Human Resource Research Center.

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 11, pp. 115–139). New York: Wiley.

*Seifert, M. K. (1995). *The relationship of role problems, work trauma, cynicism, social support, and spiritual support to the physical and mental health, work performance, and absenteeism of correctional officers* (AAT 9531303). Unpublished doctoral dissertation, University of Maryland, Baltimore County.

*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.

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.

Viswesvaran, C. (2001). Assessment of individual job performance: A review of the past century and a look ahead. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology* (Vol. 1, pp. 110–126). London: Sage.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.

*Waldman, D. A., & Avolio, B. J. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology, 76,* 897–901.

*Weekly, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52,* 679–700.

*Wing, H. (1981). Estimation of the adverse impact of a police promotion exam. *Personnel Psychology, 34,* 503–511.

Yu, C. (1998). *A study of the relationships between the self-directed learning readiness and job performance for high school principals.* Unpublished doctoral dissertation, Ohio State University, Columbus.

Received February 1, 2002
Revision received December 20, 2002
Accepted December 24, 2002 ■

Call for Nominations

The Publications and Communications (P&C) Board has opened nominations for the editorships of ***Comparative Psychology***, ***Experimental and Clinical Psychopharmacology***, ***Journal of Abnormal Psychology***, ***Journal of Counseling Psychology***, and ***JEP: Human Perception and Performance*** for the years 2006–2011. Meredith J. West, PhD, Warren K. Bickel, PhD, Timothy B. Baker, PhD, Jo-Ida C. Hansen, PhD, and David A. Rosenbaum, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2005 to prepare for issues published in 2006. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations also are encouraged.

Search chairs have been appointed as follows:

- ***Comparative Psychology***, Joseph J. Campos, PhD
- ***Experimental and Clinical Psychopharmacology***, Linda P. Spear, PhD
- ***Journal of Abnormal Psychology***, Mark Appelbaum, PhD, and David C. Funder, PhD
- ***Journal of Counseling Psychology***, Susan H. McDaniel, PhD, and William C. Howell, PhD
- ***JEP: Human Perception and Performance***, Randi C. Martin, PhD

To nominate candidates, prepare a statement of one page or less in support of each candidate. Address all nominations to the appropriate search committee at the following address:

Karen Sellman, P&C Board Search Liaison
Room 2004
American Psychological Association
750 First Street, NE
Washington, DC 20002-4242

The first review of nominations will begin December 8, 2003. The deadline for accepting nominations is **December 15, 2003.**