A Reexamination of Black–White Mean Differences in Work Performance: More Data, More Moderators

Patrick F. McKay
University of Wisconsin—Milwaukee

Michael A. McDaniel
Virginia Commonwealth University

This study is the largest meta-analysis to date of Black–White mean differences in work performance. The authors examined several moderators not addressed in previous research. Findings indicate that mean racial differences in performance favor Whites ($d = 0.27$). Effect sizes were most strongly moderated by criterion type and the cognitive loading of criteria, whereas data source and measurement level were influential moderators to a lesser extent. Greater mean differences were found for highly cognitively loaded criteria, data reported in unpublished sources, and for performance measures consisting of multiple item scales. On the basis of these findings, the authors hypothesize several potential determinants of mean racial differences in job performance.

Keywords: race, mean differences, job performance, meta-analysis, criteria

For several decades, Black–White mean differences in job performance have been a concern to personnel practitioners and researchers. This topic is of interest to practitioners because mean racial disparities in work performance may lead to differential career advancement opportunities between races (Greenhaus & Parasuraman, 1993). Furthermore, the discovery that mean racial differences in performance ratings can be attributed to bias increases organizations' vulnerability to legal scrutiny and may compromise organizational attempts to increase minority representation (Roth, Huffcutt, & Bobko, 2003).

Among researchers, mean racial differences in job performance beg the question of which aspects of performance underlie these disparities. A related concern involves whether previous research conclusions generalize to newly accumulated data. Thus far, researchers know that, on average, Whites generally are judged to perform better on the job than their Black counterparts (Chung-Yan & Cronshaw, 2002; J. K. Ford, Kraiger, & Schechtman, 1986; Hauenstein, Sinclair, Robson, Quintella, & Donovan, 2003; Kraiger & Ford, 1985; Roth et al., 2003). Standardized mean racial performance differences across these studies (d), measured in standard deviation units, range from a low of 0.24 (Chung-Yan & Cronshaw, 2002) to a high of 0.39 (Kraiger & Ford, 1985). These wide disparities in the magnitudes of Black–White mean differences in performance suggest some unexplained variability in racial effects across investigations.

The limited examination to date of potential moderators warrants reexamination of mean racial differences in work performance. For example, virtually no meta-analytic studies have assessed the influence of data source (i.e., journals vs. unpublished sources such as technical reports, dissertations, etc.) or measurement level (i.e., single-item vs. multiple-item rating scales) on mean racial differences in work performance. In addition, we are unaware of previous meta-analyses that have explored simultaneously the relative impact of multiple moderators on the magnitude of racial effects on performance. The present study meta-analyzed Black–White mean disparities in work performance with substantially more data than earlier meta-analyses.

In the sections to follow, we discuss previous meta-analytic research on overall job performance criteria and key moderators assessed in past studies. Then, several previously unstudied moderators are introduced along with justification for their consideration in the current meta-analysis.

Meta-Analytic Research on the Racial Effects on Overall Job Performance Criteria

The initial stream of meta-analytic research on Black–White mean differences in work performance focused on composite ratings of performance, without examining racial effects on the individual dimensions that comprised these ratings. Kraiger and Ford (1985) sought to clarify the magnitude of racial effects on performance and introduced the first examined moderator of Black–White mean differences in work performance, the rater–ratee race interaction. Rater–ratee race effects are operative when raters assign higher ratings to ratees of the same race. The authors reported an overall effect size for field studies of 0.39 ($k = 64$, $N = 16,149$), corrected for criterion unreliability, favoring Whites. Also, they found that Black and White raters issued higher ratings to same-race ratees (Black raters: $d = -0.45$, $k = 14$, $N = 2,428$; White raters: $d = 0.37$, $k = 74$, $N = 17,159$). It should be noted

that rater–ratee race analyses were collapsed across field and laboratory studies. Few studies involved Black raters, which prevented us from reanalyzing rater–ratee moderation here. In a second study, Waldman and Avolio (1991) examined Black–White mean performance differences using the General Aptitude Test Battery (GATB) validation study data set, and they found that Blacks received significantly lower ratings than their White counterparts. Supplemental analyses revealed that racial effects on performance were largely removed after controlling for cognitive ability, education, and job experience.

There are several reasons why the results of Kraiger and Ford (1985) and Waldman and Avolio (1991) may not be representative of current Black–White mean differences in overall job performance. First, these data preclude studies from the 1990s and beyond. Recent findings suggest that mean racial disparities in overall ratings of performance are substantially smaller than reported previously ($d = 0.27$, $k = 37$, $N = 84,295$; Roth et al., 2003). Second, the Kraiger and Ford data set contained a large number of civil service occupations, whereas Waldman and Avolio solely used the GATB database, which contains predominantly medium- to low-complexity jobs. These factors decrease the generalizability of findings. More recent data sets (Roth et al., 2003) and the current study include a broad array of occupations. Third, nearly 20 years ago, J. K. Ford et al. (1986) acknowledged that type of criterion impacts the magnitude of racial effects on performance measures. This suggests that overall performance composites may mask varying levels of racial effects on individual performance dimensions. Thus, the Kraiger and Ford and Waldman and Avolio studies offer an incomplete understanding of mean racial differences in work performance.

Additional Moderators of Racial Effects on Work Performance

A second stream of meta-analytic studies has investigated several additional moderators of Black–White mean differences in work performance. The principal moderators that have been studied include measurement method, criterion type, and the cognitive loading of criteria. During the present study, we conducted an initial examination of data source and measurement level as potential moderators of mean racial differences in job performance.

Measurement Method

Measurement method refers to whether performance criteria are measured objectively (using mechanical and/or quantified techniques) or subjectively (using ratings of performance based on judgment). It is assumed that objective performance data are less subject to racial bias than subjective evaluations of job performance such as supervisory ratings (J. K. Ford et al., 1986). Potentially, bias in performance ratings can either favor or disfavor Blacks. Raters with negative stereotypes about Blacks could deflate their ratings and perpetuate mean racial differences in performance disfavoring them (T. Cox & Nkomo, 1986; Kraiger & Ford, 1985). Accordingly, J. K. Ford et al. (1986) proposed that mean racial differences would be smaller for objective versus subjective criteria. Alternatively, raters may be motivated to inflate their ratings of Blacks to reduce the likelihood of legal scrutiny associated with low ratings (Kraiger & Ford, 1985; Mobley, 1982).

Three meta-analytic studies have examined the measurement method moderator with equivocal results. J. K. Ford et al. (1986) reported similar effect size estimates for objective ($d = 0.21$, $k = 53$, $N = 10,222$) and subjective ($d = 0.20$, $k = 53$, $N = 9,443$) criteria. However, slightly larger mean racial disparities (favoring Whites) were obtained for subjective performance indicators such as units produced and complaints ($d = 0.22$, $k = 20$, $N = 4,130$) than objective measures of these criteria ($d = 0.16$, $k = 20$, $N = 4,287$). Measurement method did moderate mean differences on cognitive criteria such as training and job knowledge, with larger effect sizes obtained for objective measures ($d = 0.34$, $k = 16$, $N = 3,389$) than subjective criteria ($d = 0.23$, $k = 16$, $N = 2,782$). Racial mean differences for absenteeism criteria varied only slightly when measured subjectively ($d = 0.15$, $k = 13$, $N = 2,221$) versus objectively ($d = 0.11$, $k = 13$, $N = 2,151$).

Chung-Yan and Cronshaw (2002) found larger mean racial differences for subjective ($d = 0.30$, $k = 57$, N not reported) versus objective ($d = 0.12$, $k = 30$, N not reported) criteria. Roth et al. (2003), on the other hand, reported larger Black–White mean differences for objective measures of quantity ($d = 0.32$, $k = 3$, $N = 774$), job knowledge ($d = 0.55$, $k = 10$, $N = 2,027$), and absenteeism ($d = 0.23$, $k = 8$, $N = 1,413$) than subjective measures of these criteria (quantity: $d = 0.09$, $k = 5$, $N = 495$; job knowledge: $d = 0.15$, $k = 4$, $N = 1,231$; absenteeism: $d = 0.13$, $k = 4$, $N = 642$). Effect sizes for quality criteria did not vary by measurement method (objective: $d = 0.24$, $k = 8$, $N = 2,538$; subjective: $d = 0.20$, $k = 10$, $N = 1,811$).

It is difficult to summarize the effects of measurement method based on the three studies reviewed above. Chung-Yan and Cronshaw (2002) showed that measurement method moderated effect sizes only for overall job performance criteria; however, J. K. Ford et al. (1986) and Roth et al. (2003) showed that method effects are confounded to some extent with criterion type (i.e., nature of performance measure). Subjective ratings resulted in larger mean racial differences in performance for noncognitive criteria and smaller mean differences for cognitive measures. A reversal of this trend occurred for objective performance indices. For comparative purposes, we revisit the issue of measurement method moderation using a larger data set than previous studies.

Criterion Type

Another research stream has explored criterion type as a moderator of racial effects on work performance. Criterion type refers to the nature of performance measures (e.g., productivity, job knowledge, and task and contextual performance). J. K. Ford et al. (1986) first addressed this issue using criterion categories that included performance indicators (e.g., units produced, accidents, and customer complaints), cognitive criteria (e.g., training and job knowledge), and absenteeism (e.g., absenteeism and tardiness). The authors failed to report overall effect sizes collapsed across subjective and objective measurement methods. Despite this concern, J. K. Ford et al. provided the first documented evidence of criterion-type moderation of Black–White mean differences in job performance. Specifically, the investigators found that cognitive criteria resulted in larger effect sizes than performance indicators and absenteeism criteria, regardless of whether performance was measured subjectively or objectively.

Three additional investigations provided insight into criterion-type effects on mean racial performance disparities (Hauenstein et al., 2003; Pulakos, White, Oppler, & Borman, 1989; Roth et al., 2003). Pulakos et al. (1989) examined Black–White mean differences in work performance for three criterion types (i.e., technical skill and effort, personal discipline, and military bearing) among army enlisted personnel. Their results showed that White soldiers received significantly higher technical skill and job effort and personal discipline ratings than Blacks. Black enlistees, in turn, earned significantly higher military bearing ratings than their White counterparts.

The Roth et al. (2003) meta-analysis investigated a number of criterion-type categories including ratings of quality (e.g., number of complaints) and quantity (e.g., number produced), job knowledge, work sample, absenteeism, on-the-job training, and promotion. Effect sizes were largest for work samples ($d = 0.52, k = 10, N = 3,651$), followed by job knowledge ($d = 0.48, k = 12, N = 2,460$), promotion ($d = 0.31, k = 7, N = 1,404$), ratings of quality ($d = 0.21, k = 15, N = 3,613$) and quantity ($d = 0.21, k = 8, N = 1,268$), absenteeism ($d = 0.19, k = 11, N = 2,376$), and on-the-job training ($d = 0.14, k = 2, N = 132$). Hauenstein et al. (2003) investigated Black–White mean differences in task performance and contextual performance. Task performance (e.g., ability to perform formal job tasks) is typically more dependent on cognitive ability than contextual performance (e.g., propensity to perform extrarole and prosocial behaviors), which is a function of personality (Borman & Motowidlo, 1993, 1997; Van Scotter & Motowidlo, 1996). This led the authors to hypothesize larger racial effects for task versus contextual performance, as supported by their subsequent findings (task: $d = 0.37, k = 10, N = 18,481$; contextual: $d = 0.27, k = 10, N = 1,634$).

During the current study, we reanalyzed the moderating effect of criterion type for task, contextual, work samples, overall job performance, on-the-job training, absenteeism, and promotion criteria using a larger number of effect sizes than previous studies. To this list of performance measures, we added academy training, personality–applied social skills, job knowledge tests, turnover, salary, accidents, and commendations–complaints.

Cognitive Loading of Criteria

Previous meta-analytic researchers have offered the post hoc hypothesis that the cognitive loading of criteria is a potential moderator of mean racial differences in work performance (J. K. Ford et al., 1986; Roth et al., 2003). To our knowledge, this hypothesis has never been formally tested. The cognitive loading of criteria connotes the degree that a criterion measure is correlated with cognitive ability (i.e., criterion-related validity). Some work-related criteria may be more dependent on cognitive ability than other criteria, which may have implications for the magnitude of racial effects. This is likely because cognitive ability is the single best predictor of job performance (Schmidt & Hunter, 1998), and White job incumbents earn significantly higher mean cognitive ability test scores than Blacks ($d = 0.90, k = 13, N = 50,799$; Roth, BeVier, Bobko, Switzer, & Tyler, 2001). In contrast, racial mean score differences on personality measures are negligible (Bobko, Roth, & Potosky, 1999; Schmitt, Clause, & Pulakos, 1996), and cognitive ability is weakly related to personality (Bobko et al., 1999). The implication of these findings is that

criteria with high cognitive loadings may result in large racial effects. In contrast, criteria that are highly correlated with personality may be less cognitively saturated and therefore inversely related to the magnitude of Black–White mean effect sizes.

Personnel theory, as well as research findings reviewed earlier, support the thesis that criterion types vary in their cognitive (and personality) loadings, which correspond to the size of racial effects on job performance. For instance, Borman and Motowidlo (1993) theorized that task performance is most dependent on cognitive ability, and contextual performance is a function of personality. Accordingly, Hauenstein et al. (2003) reported larger racial effects for task versus contextual performance measures. Similarly, Pulakos et al. (1989) found that ratings of a cognitive criterion (i.e., technical skill and job effort) favored Whites, whereas a noncognitive performance dimension (i.e., military bearing) favored Blacks. J. P. Campbell, McCloy, Oppler, and Sager (1993) reasoned that effective work sample performance requires declarative knowledge (knowledge about facts and things) and procedural knowledge (knowing what to do). The acquisition of declarative knowledge is largely dependent on general mental ability (Ackerman, 1988), why Roth et al. (2003) found large racial effects on work sample performance. Hunter (1983) theorized, and later research supported (Hunter, 1986; Schmidt, Hunter, & Outerbridge, 1986), that cognitive ability is related to job performance through the acquisition of job knowledge. The significant association between job knowledge and cognitive ability ($r = .45$; Hunter, 1986) may underlie the large Black–White mean differences in cognitive criteria and objective job knowledge criteria reported by J. K. Ford et al. (1986) and Roth et al. (2003), respectively. Other criteria, such as absenteeism, may be less influenced by cognitive ability, which explains the small racial effects obtained in prior research (J. K. Ford et al., 1986; Roth et al., 2003).

Data Source

An additional issue we examine is whether racial effects on job performance vary as a function of the source of performance data. Data source is previously unexamined as a potential moderator of mean racial differences in job performance. Publication bias has been raised as an issue in meta-analytic research (Phillips, 2004; Rothstein, 2003). Meta-analytic results will be distorted when the effect sizes in data available to researchers differ from those found in unpublished or otherwise unavailable data. In an exploratory fashion, we examine data source as a potential moderator of mean racial differences in work performance.

Measurement Level

Measurement level is another unstudied plausible moderator of Black–White mean differences in job performance. Measurement level refers to whether performance data were collected utilizing single-item or multiple-item scales. Potentially, scale-level ratings of job performance constructed from single-item ratings of multiple performance dimensions are more reliable than single-item ratings taken in isolation. Unreliability in single-item criteria may underestimate population-level racial effects on performance ratings (Hunter & Schmidt, 1990, 2004). Although measurement level is not a theoretically meaningful moderator, we felt it worth-

while to examine its potential effects on Black–White mean differences in work performance.

The Present Study

The current investigation further assesses the magnitude of Black–White mean differences in work performance. We feel that this study adds to the personnel literature in a number of ways. First, our study represents the largest and most comprehensive meta-analysis to date of mean racial differences in work performance. Second, these additional data enable us to provide more stable effect size estimates for criterion-type categories whose results have been based on few studies in past research (e.g., work samples, task performance, and contextual performance). Third, our large data set also allowed us to analyze interactions among multiple moderators. Last, we consider several unexamined moderators that may influence mean racial differences in work performance.

Method

Sample of Studies

We used a series of methods to gather studies appropriate for inclusion in our meta-analysis. First, we searched the PsycINFO, ABI/Inform, Social Sciences Index, Educational Resources Information Center, and Dissertation Abstracts databases for articles and theses–dissertations that may contain useful data. Searches were performed using the keywords *job performance*, *performance ratings*, *performance evaluation*, *criteria*, *validation*, and combination of keywords (e.g., *racial differences* and *job performance*, *performance ratings* and *race*). Second, we consulted previous meta-analyses to identify studies for inclusion (e.g., Chung-Yan & Cronshaw, 2002; J. K. Ford et al., 1986; Kraiger & Ford, 1985; Roth et al., 2003). Third, performance appraisal and test validation researchers–practitioners were solicited to provide performance data from unpublished data sources. Lastly, we performed manual searches of industrial–organizational psychology, organizational behavior, and testing-oriented research journals (e.g., *Academy of Management Journal*, *Educational and Psychological Measurement*, *Journal of Applied Psychology*, and *Personnel Psychology*). Using these methods, we identified a total of 146 studies.

Criteria for Inclusion

Eight decision rules were used to evaluate studies for inclusion in the meta-analysis. First, studies were included if they provided supervisory ratings of job performance and/or performance data from personnel records with means, standard deviations, subgroup sample sizes or other statistics such as F tests that enabled the computation of effect sizes. Several investigations were eliminated for failing to provide necessary statistical indices (e.g., C. L. Moore, MacNaughton, & Osburn, 1969; Thompson & Thompson, 1985; Vosburgh, 1988).

Second, data were only extracted from studies that contained incumbent work performance, turnover records, and/or academy training performance. Studies that used (a) job applicants (e.g., Cascio & Phillips, 1979), (b) college student samples (e.g., Bigoness, 1976; Feldman & Hilterman, 1977; Hall & Hall, 1976; Hamner, Kim, Baird, & Bigoness, 1974; Outtz, 1977; Rotter & Rotter, 1969; Schmitt & Lappin, 1980), (c) ratings generated from videotaped performance (Brugnoli, Campion, & Basen, 1979), and (d) upward, peer, or self-appraisals (e.g., J. A. Cox & Krumboltz, 1958; deJung & Kaplan, 1962; Grant-Vallone, 1998; Schmidt & Johnson, 1973) were not included in the meta-analysis.

Third, studies were selected if employee performance data were reported separately for Black and White subgroups. Therefore, investigations that collapsed performance ratings across several minority groups were excluded (e.g., Cascio & Valenzi, 1978; Feild, Bayley, & Bayley, 1977; Kesselman & Lopez, 1979; Morstain, 1984; Toole, Gavin, Murdy, & Sells, 1972).

Fourth, we only analyzed performance data used for criterion-related validation studies and/or administrative, personnel decision-making purposes. Studies that collected performance data for developmental purposes were discarded (e.g., Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Mount, Sytsma, Hazucha, & Holt, 1997; Tuzinski & Ones, 1999).

Fifth, if data were reported in two sources, we only used the data from one source. Using this decision rule necessitated the exclusion of several data sets. For example, Sample 1 data from Clevenger, Pereira, Wiechmann, Schmitt, and Harvey (2001) were also used in Pulakos and Schmitt (1996). As a result, Sample 1 from Clevenger et al. (2001) was not included in meta-analytic estimates. Similarly, we excluded data from Pulakos, Schmitt, and Chan (1996) and Pulakos and Schmitt (1995) because each of these investigations contained data reported in Pulakos and Schmitt (1996).

Sixth, similar to Roth et al. (2003), we eliminated studies subject to range enhancement. In such cases, criterion data were collected in a way that artificially expands the range of performance. Data from one study (Baehr, Saunders, Froemel, & Furcon, 1971) were excluded because the researchers purposely sampled police incumbents from the upper and lower thirds of performance distributions, which would lead to overestimates of racial mean effect sizes.

Seventh, we only included data from studies in which Black and White employees worked in the same organization, as suggested by Roth et al. (2003). Failure to satisfy this decision rule required us to eliminate data from Study 4 of Kirkpatrick, Ewen, Barrett, and Katzell (1968).

Lastly, we eliminated criterion data if they failed to describe actual work and/or training performance. Thus, days of vacation and education and training score criteria from Neidt (1968, Samples 1 and 3) and garnishments per year (Samples 3 and 4) were omitted. Unlike Roth et al. (2003), we did not eliminate studies that obtained data from organizations that adhered to affirmative action programs or studies that collected data from firms in which racial groups differed in their opportunity to perform job duties (e.g., Black and White police officers assigned to policing sectors with differential crime rates). Specifically, we analyzed data from Bartlett et al. (1977), Kraiger (1981), and Mills (1990) that were excluded from the Roth et al. (2003) study. In our view, the inclusion of these studies will increase the generalizability of our results. Furthermore, the presence of affirmative action programs and/or differential opportunity to perform in firms does not indicate biased performance ratings per se.

Using the above selection criteria, we judged 97 studies as acceptable for inclusion in the meta-analysis. Our 97 studies exceed the 36 studies included in Roth et al.'s (2003) recent study. Comparison of the studies included in our meta-analysis with Roth et al.'s (2003) study showed that the two meta-analyses shared 28 studies, whereas Roth et al. (2003) included eight studies we were unable to acquire. Barring these eight studies, our meta-analysis contains data from an additional 61 studies not included in the Roth et al. (2003) meta-analysis. This is an important strength of the current investigation, because our substantial additional data allow for more precise estimation of racial effects on job performance. A significant source of the sample size difference between the two meta-analyses is that we gathered a greater number of unpublished data sources than Roth et al. (2003). A potential implication is that the estimates reported, for at least some criterion-type distributions, may vary from those reported by Roth et al. (2003), because publication bias distorts effect sizes in meta-analytic research (Rothstein, 2003). Most of our unpublished data were derived from the GATB data set and a number of consulting firms, particularly Hogan Assessment Systems. In addition, we examined three moderators not considered by Roth et al. (2003), namely the cognitive loading of criteria, data source, and measurement level. On the basis of these factors, the present study is most comparable with Roth et al. (2003)

for assessing the magnitude of effect sizes from published articles and conference presentations.

Data Coding

Each study was read to search for moderators and other characteristics of interest to our investigation. The coding schemes used to categorize moderators and study characteristics are explained below.

Criterion type. Using both the job performance taxonomy developed by Borman and Motowidlo (1993) and Roth et al.'s (2003) 11-category scheme, two independent coders (one a business management doctoral student and the other a master's level industrial–organizational psychologist) assigned each performance measure to 1 of 12 job performance categories listed in Table 1 or to measures of academy training performance or turnover. An intercoder agreement of 92% was obtained, and Patrick F. McKay resolved any coding discrepancies.

Our criterion-type coding scheme differed from Roth et al.'s (2003) scheme in several ways. First, we subsumed ratings of quality, quantity, and job knowledge within the task criterion category when performance data represented core tasks associated with a given job. We felt this decision was justified for several reasons. Measures of quantity of output were of low frequency in our data set, which would reduce the stability of effect size estimates derived for these criteria as a separate category. In addition, quantity criteria represent core performance in such jobs as manufacturing and sales, meaning that these indices are reflective of task performance. Also, Roth et al. (2003) obtained identical effect sizes for quantity and quality criteria ($d = 0.21$), and their effect size distributions overlapped considerably ($0.14–0.27$ for quality and $0.03–0.40$ for quantity).

Ratings of job knowledge were classified as task criteria because these measures were defined in source documents as evaluations of incumbents' effective use of job knowledge in completing job tasks, and not their extent of knowledge per se. In our view, this conception of job knowledge is more representative of task proficiency as defined by Borman and Motowidlo (1993) and qualitatively different than the extent of knowledge as measured by objective job knowledge tests. Roth et al. (2003) found that subjective measures of knowledge resulted in smaller effect sizes than objective measures (subjective $d = 0.15$, objective $d = 0.55$), which

supports our rationale for treating ratings of job knowledge separately. An implication of this decision is that effect sizes for task criteria may be elevated to the extent that ratings of job knowledge are cognitively loaded.

A second way that our classification scheme differed from Roth et al.'s (2003) is that organizational citizenship behaviors and other established measures of contextual performance (e.g., interpersonal facilitation) were coded using definitions provided by Borman and Motowidlo (1993). Third, criteria that did not sort neatly into the contextual performance category were classified as personality–applied social skills measures, because these criteria were ratings of incumbents' work-related personality traits or interpersonal skills (e.g., even tempered, gets along with fellow employees). This decision was based on whether the performance measures were developed as indicators of contextual performance as described in source studies. Huffcutt, Conway, Roth, and Stone (2001) stated that interview ratings of personality constructs and applied social skills are associated such that social skills reflect inherent personality tendencies. Furthermore, the authors reported that personality and applied social skills resulted in similar levels of racial effects on interview ratings (e.g., extroversion: $d = 0.18$, $k = 3$, $N = 1,333$; interpersonal skills: $d = 0.22$, $k = 6$, $N = 1,733$), so we felt justified in placing these criteria in the same category.

Fourth, we sorted job knowledge tests into a unique category. We believe that this classification corresponds to Roth et al.'s (2003) job knowledge–objective criterion category. It should be noted that job knowledge tests used in the evaluation of training academy performance were included here as well. Fifth, training measures were subdivided into (a) academy training criteria based on the scholastic learning of job duties in a classroom setting (e.g., final academy grade) and assumedly more dependent on cognitive ability and (b) on-the-job training conducted on the job site with lower cognitive requirements (e.g., final field performance rating). Sixth, absenteeism–lost time criteria included both attendance and lateness data given their conceptual overlap as measures of employee work withdrawal behaviors (Blau, 1994; Koslowsky, Sagie, Krausz, & Singer, 1997). Finally, salary was considered a separate criterion domain from promotion because salary increases do not require promotion, whereas the obverse is usually true.

Measurement level. For each performance measure, data were coded to distinguish single-item dimension ratings from composite, scale-level ratings comprised of multiple job performance dimensions that are summed

Table 1
Criterion-Type Categories

Category	Work performance criteria
Task performance	Ratings of proficiency in performing core duties of a position
Contextual performance	Ratings of performance of prosocial and extrarole behaviors beyond assigned work duties
Personality–applied social skills	Ratings of stable behavioral tendencies and/or social skills related to behaving effectively in work-related social situations (e.g., even tempered, gets along with fellow employees)
On-the-job training	Measures of training effectiveness collected on the job (e.g., ability to learn)
Overall job performance	Summary ratings of overall effectiveness on the job
Work sample	Scores earned on tests designed to simulate aspects of work task
Job knowledge test	Tests of training mastery or knowledge (includes job knowledge tests from training academies)
Attendance–lost time	Objective or subjective measures of work attendance and tardiness
Promotion	Objective or subjective measures of promotions or transfers to other positions
Salary	Objective measures of salary, merit increases, etc.
Accidents	Objective or subjective measures of the number of accidents in which an employee is involved
Commendations–reprimands	Objective or subjective measures of violations or awards received, in addition to criteria referring to counterproductive work behaviors (e.g., service complaints)

or averaged to derive criteria. Intercoder agreement for measurement level coding was 94%. Again, Patrick F. McKay resolved disagreements.

Cognitive loading of criteria. For each sample, the cognitive loading of criteria was defined as the Pearson product–moment correlation coefficient between cognitive ability test scores and work performance criteria. Not all samples that provided a mean racial difference effect size contained cognitive loading of criteria data. Thus, these samples were not included in cognitive load analyses.

Data source. Data source was coded by Patrick F. McKay and represents the type of reference source from which data were extracted. Sources of reported data included journal articles, doctoral dissertations, conference papers, and unpublished technical reports. We also received some primary study data sets used to calculate effect sizes, such as data from 34,219 individuals who participated in the GATB validation studies. Results of these analyses were classified as unpublished technical reports because of their lack of public availability.

Measurement method. Performance data were coded as either gathered from objective, quantified measures of performance (e.g., number of absences) or subjective supervisory ratings of performance. The coders agreed 99% on measurement method codes, with discrepancies resolved by Patrick F. McKay.

Calculation of Effect Sizes

We conducted a psychometric meta-analysis of standardized mean differences (Hunter & Schmidt, 1990, 2004). Analyses were performed in SAS using code adapted from Arthur, Bennett, and Huffcutt (2001). We verified the accuracy of our analyses by comparing the program output with that produced by Schmidt and Le's (2004) psychometric meta-analysis software. For the effect size estimates (d) reported, positive ds mean that Whites' performance exceeds Blacks', whereas negative ds mean that Blacks outperformed Whites.

In our meta-analysis, we focused on uncorrected effect sizes to provide conservative conclusions regarding the magnitude of Black–White mean differences in work performance; however, data tables also include effect sizes corrected for measurement error to allow comparability with previous meta-analytic studies. The procedures used to compute these corrected effect size estimates are explained briefly. Roth et al. (2003) assumed reliabilities .8 and .6 for objective and subjective criteria, respectively; therefore, we used these values for objective and subjective criteria measured at the scale level. Because item-level reliabilities rarely exceed .25 (Hunter & Schmidt, 1990), we estimated the reliability of single-item measures, assuming that scale-level measures were composed of five items. This assumption is conservative because most performance scales contained more than five items. Then, using scale-level reliability estimates, we applied the Spearman–Brown formula to estimate reliability for single-item subjective and objective measures, and we corrected effect sizes using the resulting upper-bound reliability estimates of .23 and .44, respectively.

To examine the cognitive loading moderator, we conducted correlated vectors meta-analyses (Jensen, 1998). These analyses computed the correlation between mean criterion cognitive loading and mean criterion racial difference effect size vectors, weighted by sample size, in which each sample is the unit of analysis. A positive correlation indicates that the size of mean racial differences in performance increases as the cognitive loading of criteria increases. To compare with cognitive loading results, we conducted correlated vectors analysis examining the correlations between personality dimensions and criteria (i.e., personality loading of criteria) and their relationship with mean criterion racial difference effect sizes. The six personality factors examined were conscientiousness (i.e., degree that one is dependable or disciplined), emotional stability (i.e., extent that a person is self-confident and calm), agreeableness (i.e., degree that one is sensitive to others), ambition (i.e., degree that an individual is hardworking), openness (i.e., degree that one is curious and eager to learn), and school success (i.e., extent that a person has high achievement orientation; J. Hogan &

Holland, 2003). We note that all personality loading of criteria data came from unpublished technical reports from Hogan Assessment Systems. Although many employers contributed data to this personality data set, analyses conducted by a single consulting firm are likely to involve similar procedures and assessment devices. As a consequence, the data gathered might be less variable than data collected from multiple sources.

In addition, we assessed the correlations between each of the five moderators and effect sizes. Because criterion type, measurement level, measurement method, and data source are nominal variables with multilevel classifications, their relations with effect sizes are reported as multiple correlations. The multiple correlation for criterion type was corrected for shrinkage using the formula provided by Cohen and Cohen (1983). Cognitive loading and personality loading are continuous measures, so their associations with effect sizes represent bivariate correlations. Many studies did not report cognitive ability or personality validity data; therefore, cognitive loading and personality loading results are available for only a subset of the data. Furthermore, no studies contained both cognitive ability and personality measures, which prevented us from directly comparing cognitive load and personality load moderating effects within studies.

Effect Sizes per Sample

Effect size estimates are presented such that one effect size is reported for each sample. Most studies included criteria from only one type category, whereas several investigations contained data from multiple categories (e.g., Johnson, 2001). In these instances, we included multiple effect sizes that corresponded to each criterion-type category. In those rare samples in which a given criterion type contained both subjective and objective criteria, we estimated effect sizes separately for each measurement method. A few studies included a single criterion type (e.g., contextual performance) measured at both the single-item and scale levels (e.g., Huck & Bray, 1976; Johnson, 2001). To examine measurement level moderation, we included two effect sizes to summarize both single-item and scale-level measurement effects.

Results

We focus our reporting on uncorrected effect sizes as conservative estimates of mean racial differences in work performance. Where applicable (i.e., academy training results), separate mean effect size estimates are provided for analyses that incorporated or excluded large sample studies. Investigations with samples that exceeded 2,000 employees were identified as large sample studies, in line with Roth et al.'s (2003) recommendation. Also, results are categorized by those including and excluding effect sizes from the GATB data set. Because these data include a preponderance of medium- and low-complexity jobs, analyses containing these data may influence effect size magnitudes. The Department of Labor used the same or similar procedures to collect GATB data from many private employers that hired workers registered with the United States Employment Service. Accordingly, these data may provide better control over study-specific extraneous variation than the rest of our data. Also, because we had all "individual observation" GATB data as of August 2003, effect sizes that we calculated from this data set were not likely affected by publication bias. For these reasons, whether effects were drawn from the GATB database may serve as a moderator.

Black–White Mean Differences in Job Performance

Table 2 presents results pertaining to Black–White mean differences in job performance summarized across criteria, excluding

Table 2
Black–White Mean Differences in Job Performance × Measurement Level, Measurement Method, Data Source, and Criterion Type

Distribution analyzed	d	k	N_{Total}	N_{White}	N_{Black}	90% CI	PVA	$d_{\text{corrected}}$
All ratings	0.27	572	109,974	79,092	30,882	−17 to 0.71	23	0.38
GATB studies	0.40	179	34,219	25,146	9,073	0.07 to 0.74	34	0.52
No GATB	0.21	393	75,755	53,946	21,809	−24 to 0.66	22	0.31
Military studies	−0.09	5	4,067	2,817	1,250	−0.45 to 0.26	10	−0.12
No military	0.28	567	105,907	76,275	29,632	−0.15 to 0.71	24	0.40
Measurement level								
Single-item ratings	0.15	187	36,939	24,582	12,357	−0.22 to 0.51	29	0.29
Scale ratings	0.33	385	73,035	54,510	18,525	−0.11 to 0.77	23	0.42
GATB studies	0.40	178	34,115	25,084	9,031	0.07 to 0.74	34	0.52
No GATB	0.27	207	38,920	29,426	9,494	−0.23 to 0.77	19	0.34
Measurement method								
Subjective	0.28	510	94,555	69,271	25,284	−0.15 to 0.70	25	0.40
GATB studies	0.40	176	33,868	24,877	8,991	0.07 to 0.74	34	0.52
No GATB	0.21	334	60,687	44,394	16,293	−0.22 to 0.64	25	0.33
Objective	0.22	62	15,419	9,821	5,598	−0.31 to 0.74	14	0.27
GATB studies	0.28	3	351	269	82	0.10 to 0.47	73	0.31
No GATB	0.21	59	15,068	9,552	5,516	−0.31 to 0.74	14	0.27
Data source								
Dissertation	0.12	36	6,291	3,462	2,829	−0.09 to 0.34	57	0.20
Journal	0.17	118	32,026	21,545	10,481	−0.30 to 0.64	15	0.24
Conference paper	0.38	16	2,330	1,398	932	0.14 to 0.62	58	0.71
Technical report	0.34	394	67,646	51,409	16,237	−0.07 to 0.75	28	0.46
GATB studies	0.40	179	34,219	25,146	9,073	0.07 to 0.74	34	0.52
All unpublished data, no GATB	0.27	215	33,427	26,263	7,164	−0.18 to 0.72	26	0.39
Book	−0.01	8	1,681	1,278	403	−0.24 to 0.21	52	−0.05
Criterion-type category								
Task	0.21	93	15,868	10,901	4,967	−0.07 to 0.50	44	0.35
Contextual	0.13	31	3,333	2,491	842	−0.24 to 0.51	42	0.21
Personality–applied social skills	0.07	60	10,648	7,250	3,398	−0.33 to 0.47	28	0.15
On-the-job training	0.05	7	1,510	1,022	488	−1.30 to 1.40	3	0.25
Overall job performance	0.35	302	58,808	44,906	13,902	−0.01 to 0.71	31	0.46
GATB studies	0.40	176	33,868	24,877	8,991	0.07 to 0.74	34	0.52
No GATB	0.28	126	24,940	20,029	4,911	−0.08 to 0.63	31	0.39
Work sample	0.42	23	6,557	4,201	2,356	0.12 to 0.71	31	0.56
GATB studies	0.28	3	351	269	82	0.10 to 0.47	73	0.31
No GATB	0.42	20	6,206	3,932	2,274	0.13 to 0.72	29	0.57
Job knowledge test	0.53	9	2,216	1,360	856	0.33 to 0.74	53	0.60
Absenteeism–lost time	0.09	20	3,779	2,218	1,561	−0.39 to 0.56	20	0.13
Salary	0.14	5	1,233	919	314	−0.25 to 0.53	22	0.21
Promotion	0.18	7	1,422	1,152	270	−0.11 to 0.47	39	0.34
Accidents	−0.06	6	2,371	1,275	1,096	−0.41 to 0.29	18	−0.09
Commendations–reprimands	0.02	9	2,229	1,397	832	−0.49 to 0.52	15	0.02

Note. All General Aptitude Test Battery (GATB) data were measured at the scale level. CI = confidence interval; PVA = percentage of variance accounted for by sampling error.

academy training and turnover measures. The mean Black–White difference in performance is 0.27 favoring Whites, which is identical to the estimate reported by Roth et al. (2003). This summary effect size is nearly twice as large for GATB studies ($d = 0.40$, $k = 179$) than no-GATB studies ($d = 0.21$, $k = 393$); however, GATB studies only included work performance data measured at the scale level, whereas no-GATB studies contained a mixture of single-item and scale-level criteria. A more appropriate comparison of GATB and no-GATB studies is presented for scale-level overall job performance criteria results (see Table 7), which shows that the corresponding effect sizes are similar in magnitude (ds = 0.40 and 0.35, respectively). As shown in Table 2, military studies resulted in substantially smaller racial effects on performance ($d = -0.09$, $k = 5$) than no-military studies ($d = 0.28$, $k = 567$).

Moderator Analyses

We proposed that five moderators would influence mean racial differences in work performance, including criterion type, cogni-

tive loading of criteria, data source, measurement level, and measurement method. The results for academy training and turnover criteria are presented first, followed by findings that summarize the relative influence of each moderator on the size of racial effects on job performance. Lastly, findings for each moderator are briefly presented.

The results for academy training and turnover are displayed in Tables 3 and 4, respectively. Table 5 presents summary moderator results for job performance criteria, with more detailed evidence about each moderator provided in Tables 2, 3, 6, and 7. As

reported in Table 3, academy training measures resulted in substantial Black–White mean differences ($d = 0.46$). Whites obtained training proficiency scores–ratings that are nearly one half standard deviation higher than Blacks. Eliminating large sample studies ($k = 2$) from analyses results in an even larger effect size ($d = 0.65$, $k = 23$). Academy training effect sizes varied on the basis of whether data were collected in military ($d = 0.42$, $k = 3$) versus nonmilitary or civilian ($d = 0.66$, $k = 22$) settings. In addition, we found that mean racial differences in academy training proficiency were smaller for non-GATB ($d = $

Table 3

Black–White Mean Differences in Academy Training Performance $\times$ Measurement Level and Measurement Method

Criterion	d	k	N_{Total}	N_{White}	N_{Black}	90% CI	PVA	$d_{\text{corrected}}$
All training criteria	0.46	25	31,307	29,137	2,170	0.15 to 0.77	9	0.72
GATB studies	0.69	3	724	575	149	0.16 to 1.23	14	0.78
No GATB	0.45	22	30,583	28,562	2,021	0.16 to 0.75	9	0.72
Large sample studies	0.42	2	25,928	24,906	1,022	0.31 to 0.53	6	0.71
No large sample	0.65	23	5,379	4,231	1,148	0.05 to 1.25	12	0.81
Military studies	0.42	3	26,115	25,008	1,107	0.31 to 0.53	9	0.70
No military	0.66	22	5,192	4,129	1,063	0.06 to 1.27	12	0.82
Measurement level								
Single-item ratings	0.49	7	13,643	12,857	786	0.40 to 0.58	42	1.01
Large sample studies	0.49	1	12,964	12,453	511			1.02
No large sample	0.45	6	679	404	275	0.04 to 0.86	37	0.72
Scale ratings	0.44	18	17,664	16,280	1,384	0.04 to 0.83	7	0.50
GATB studies	0.69	3	724	575	149	0.16 to 1.23	14	0.78
No GATB	0.43	15	16,940	15,705	1,235	0.04 to 0.81	6	0.49
Large sample studies	0.35	1	12,964	12,453	511			0.39
No large sample	0.68	17	4,700	3,827	873	0.07 to 1.29	10	0.82
Measurement method								
Subjective	0.51	11	15,508	14,555	953	0.32 to 0.70	17	0.99
Large sample studies	0.49	1	12,964	12,453	511			1.02
No large sample	0.61	10	2,544	2,102	442	0.16 to 1.06	18	0.80
Objective	0.41	14	15,799	14,582	1,271	0.04 to 0.78	7	0.47
GATB studies	0.69	3	724	575	149	0.16 to 1.23	14	0.78
No GATB	0.40	11	15,075	14,007	1,068	0.05 to 0.74	6	0.45
Large sample studies	0.35	1	12,964	12,453	511			0.39
No large sample	0.69	13	2,835	2,129	706	−0.01 to 1.40	10	0.81
Data source								
Journal	0.42	2	25,928	24,906	1,022	0.31 to 0.53	6	0.71
Conference paper	0.47	2	291	149	142	0.47 to 0.47	100	0.65
Technical report	0.66	21	5,088	4,082	1,006	0.05 to 1.28	11	0.81
GATB studies	0.69	3	724	575	149	0.16 to 1.23	14	0.78
No GATB	0.66	18	4,364	3,507	857	0.03 to 1.29	11	0.82
Data Source $\times$ Measurement Method $\times$ Measurement Level								
Single-item rating								
Subjective	0.49	3	13,101	12,559	542	0.49 to 0.49	100	1.02
Objective	0.47	4	542	298	244	−0.02 to 0.97	25	0.71
Scale-level rating								
Subjective	0.62	8	2,407	1,996	411	0.16 to 1.08	15	0.80
Conference paper	0.63	1	134	71	63			0.81
Technical report	0.62	7	2,273	1,925	348	0.14 to 1.10	13	0.80
Objective	0.41	10	15,257	14,284	973	0.05 to 0.77	5	0.46
Journal	0.35	1	12,964	12,453	511			0.39
Technical report	0.74	9	2,293	1,831	462	0.02 to 1.46	8	0.83

Note. All General Aptitude Test Battery (GATB) data were measured at the scale level. CI = confidence interval; PVA = percentage of variance accounted for by sampling error.

Table 4
Black–White Mean Differences in Turnover

Distribution analyzed	d	k	N_{Total}	N_{White}	N_{Black}	90% CI	PVA	$d_{corrected}$
Turnover	−0.30	6	1,336	856	480	−0.92 to 0.31	12	−0.46

Note. All data are single-item ratings and objective measurement method. CI = confidence interval; PVA = percentage of variance accounted for by sampling error.

0.43, $k = 15$) than GATB ($d = 0.69$, $k = 3$) data measured at the scale level. Lastly, measurement method appeared to moderate racial effects on academy training performance as effect sizes were larger overall for subjective ($d = 0.51$) versus objective ($d = 0.41$) criteria and when measured at the scale level (subjective $d = 0.62$, objective $d = 0.41$). This effect did not hold for single-item academy training measures (subjective $d = 0.49$, objective $d = 0.47$).

The turnover criteria effect size of −0.30 suggests that Blacks turn over less than Whites (see Table 4). This estimate should be

Table 5
Correlated Vectors Analyses of the Strength of Moderators in Accounting for Black–White Mean Differences in Job Performance

Moderator	R (r)	k
Measurement level	.28	572
Measurement method	.10	572
Criterion type	.40	572
Data source	.30	572
Criterion cognitive loading	.34	291
Task criteria	.13	31
Item level	.06	21
Scale level	.42	10
Contextual	.41	18
Item level	−.12	9
Scale level	.64	9
Personality–applied social skills	.20	5
Item level	−.67	4
Overall performance criteria	.32	202
Item level	.50	11
Scale level	.31	191
Work sample	.90	13
Item level	.91	4
Scale level	.89	9
Job knowledge test	.71	6
Absenteeism–lost time	−.17	5
Promotion	.77	4
5 nonpersonality moderator set	.62	291
Conscientiousness	−.23	138
Emotional stability	−.46	90
Agreeableness	−.06	135
Ambition	−.17	90
Openness	−.11	96
School success	−.17	96
6 personality moderator set	.56	90
6 personality moderators plus other moderators (excluding cognitive loading)[a]	.58	90

[a] Given the more homogeneous study characteristics found in the data from Hogan Assessment Systems, the source for most of the personality data, measurement level, measurement method, and data source were constants in this analysis. This restricts the multiple R to a level lower than it would be if the variance of the three variables were not zero.

interpreted cautiously because it is based on few effect sizes ($k = 6$). Because all turnover criteria were involuntary termination measures, it appears that organizations, on average, are less likely to fire Black employees.

Table 6 presents correlated vectors meta-analytic findings that summarize the relative strength of relationship between the five moderators and mean racial job performance difference effect sizes. Our findings show that Black–White mean differences in work performance are most strongly moderated by criterion type ($R = .40$), followed in order by the cognitive loading of criteria ($r = .34$), data source ($R = .30$), measurement level ($R = .28$), and measurement method ($R = .10$). Overall, the five moderators accounted for 38% of the variance in mean racial differences in job performance ($R = .62$). Although our correlational analyses provide a useful overview of moderator strength, Tables 2, 3, 6, and 7 present the magnitude of mean racial effect size differences within moderator and nested moderator groups.

Table 2 shows that mean racial differences in job performance are greatest for job performance measures highly dependent on cognitive ability (Hunter, 1983; McCloy, Campbell, & Cudeck, 1994) or its correlate, declarative knowledge (Ackerman, 1988), which includes work samples and job knowledge tests ($ds = 0.42$ and 0.53, respectively). This result did not hold for task performance criteria ($d = 0.21$). Our estimate for job knowledge tests is similar to the objectively measured job knowledge estimate reported by Roth et al. (2003; $d = 0.55$, $k = 10$, $N = 2,027$), whereas the effect size for work samples is somewhat lower than the value obtained by those authors ($d = 0.52$, $k = 10$, $N = 3,651$). The discrepancy between work sample effect sizes is likely the result of our analysis of a larger number of effect sizes ($k = 23$). Seemingly less cognitively loaded measures such as contextual ($d = 0.13$), personality–applied social skills ($d = 0.07$), and absenteeism–lost time ($d = 0.09$) criteria, as well as overall job performance ($d = 0.35$), resulted in smaller racial effects on performance. Compared with Roth et al. (2003; $d = 0.19$, $k = 11$, $N = 2,376$), our effect size for absenteeism–lost time criteria is smaller in magnitude, perhaps because of our inclusion of absence and lateness data in this category and a larger number of effect sizes ($k = 20$). The relationship between criterion type and cognitive load may explain variability in the magnitudes of effect sizes across various criterion categories. Table 8 reproduces mean racial difference effect sizes from Table 2 and includes the mean cognitive loading value for each criterion. Correlated vector analyses revealed that the mean effect size and mean cognitive load vectors are highly related ($r = .66$). The correlation drops to .60 after weighting it by the number of studies (k) analyzed.

In contrast, findings for personality loading run counter to the cognitive loading results, as expected. All six personality loading–effect size vector correlations are negative, which indicates that

Table 6
Black–White Mean Differences in Job Performance for Single-Item Data: Criterion Type × Measurement Method × Data Source

Distribution analyzed	d	k	N_{Total}	N_{White}	N_{Black}	90% CI	PVA	$d_{\text{corrected}}$
Task	0.19	56	9,986	6,812	3,174	−0.08 to 0.45	47	0.37
Subjective	0.18	48	8,263	5,596	2,667	0.03 to 0.34	72	0.38
Dissertation	0.09	7	1,193	698	495	0.09 to 0.09	100	0.19
Journal	0.17	16	2,856	1,974	882	−0.14 to 0.48	39	0.35
Conference paper	0.31	4	592	333	259	0.20 to 0.42	86	0.64
Technical report	0.21	20	3,552	2,552	1,000	0.21 to 0.21	100	0.43
Book	0.11	1	70	39	31			0.24
Objective	0.20	8	1,723	1,216	507	−0.35 to 0.74	15	0.30
Contextual (all from journals and subjective data source)	0.12	9	1,240	890	350	−0.06 to 0.29	72	0.24
Personality–applied social skills (all subjective)	0.14	31	5,893	3,914	1,979	−0.03 to 0.30	68	0.28
Dissertation	0.10	6	1,028	615	413	0.10 to 0.10	100	0.20
Journal	0.05	6	1,307	804	503	−0.19 to 0.28	48	0.10
Conference paper	0.35	4	592	333	259	0.11 to 0.58	58	0.72
Technical report	0.15	15	2,966	2,162	804	0.15 to 0.15	100	0.31
On-the-job training	0.45	4	1,031	681	350	0.12 to 0.78	29	0.86
Overall job performance (all subjective)	0.15	35	6,606	4,666	1,940	−0.20 to 0.50	31	0.31
Dissertation	0.23	3	659	398	261	0.10 to 0.35	75	0.47
Journal	0.14	6	763	494	269	0.14 to 0.14	100	0.29
Conference paper	0.36	5	725	405	320	0.13 to 0.58	60	0.74
Technical report	0.14	18	3,717	2,794	923	−0.23 to 0.52	28	0.30
Book	−0.09	3	742	575	167	−0.44 to 0.26	27	−0.19
Work sample	0.44	7	1,343	803	540	0.32 to 0.57	79	0.80
Subjective	0.52	3	576	344	232	0.52 to 0.52	100	1.09
Objective	0.39	4	767	459	308	0.21 to 0.56	66	0.58
Absenteeism–lost time	0.07	18	3,585	2,073	1,512	−0.40 to 0.53	20	0.10
Subjective	−0.01	6	1,245	584	661	−0.17 to 0.15	67	−0.02
Objective	0.11	12	2,340	1,489	851	−0.45 to 0.66	15	0.16
Dissertation	0.12	2	272	170	102	0.12 to 0.12	100	0.17
Journal	0.18	7	1,748	1097	651	−0.11 to 0.47	35	0.27
Technical report	−0.31	3	320	222	98	−1.45 to 0.84	7	−0.46
Salary (all objective)	0.14	5	1,233	919	314	−0.25 to 0.53	22	0.21
Promotion	0.18	7	1,422	1,152	270	−0.11 to 0.47	39	0.34
Accidents	−0.06	6	2,371	1,275	1,096	−0.41 to 0.29	18	−0.09
Commendations–reprimands	0.02	9	2,229	1,397	832	−0.49 to 0.52	15	0.02

Note. All General Aptitude Test Battery (GATB) data were measured at the scale level. CI = confidence interval; PVA = percentage of variance accounted for by sampling error.

highly personality saturated criteria are associated with reduced Black–White mean differences in job performance. Emotional stability is most strongly correlated with effect sizes ($r = -.46$), followed by conscientiousness ($r = -.23$), ambition ($r = -.17$), school success ($r = -.17$), and agreeableness ($r = -.06$). The reversed pattern of findings between cognitive loading and personality loading of criteria is suggestive that the cognitive loading and personality loading of criteria may be inversely related. In other words, performance measures with high correlations with cognitive ability have low, near-zero, positive correlations with personality. Overall, the six personality variables accounted for 31% of variance in effect sizes ($R = .56$).

Data source was explored as a potential moderator of mean racial differences in work performance. Data source moderation would be evident if journal outlets show lower magnitude effect sizes than unpublished sources such as dissertations, conference papers, and technical reports. Our findings provide some support for this proposition. Table 2 shows that mean racial job performance effect sizes, summarized across criteria, are greater for conference papers ($d = 0.38$) and technical reports ($d = 0.34$; no GATB $d = 0.27$) than journals ($d = 0.17$), dissertations ($d =$

0.12), and books ($d = -0.01$). Data source effects appear to vary on the basis of criterion type and measurement level, as shown in Tables 3 and 7.

Measurement level refers to whether work performance data were gathered from single-item or multiple-item scales, assuming that the latter are more reliable than the former. Accordingly, we expected Black–White mean differences in job performance to be larger for scale-level versus single-item criteria. Measurement level results are presented in Tables 2, 3, 6, and 7. As reported in Table 2, findings for measurement level, collapsed across all job performance criteria, suggest some evidence of moderation. Mean racial differences in job performance are over twice as large for criteria measured by scales ($d = 0.33$) than with single items ($d = 0.15$). The direction and magnitude of measurement level effects on mean racial differences in performance become less clear when considered for specific criterion-type categories (see Tables 3, 6, and 7). Support for measurement-level moderation is clearly evident for overall job performance (single-item $d = 0.15$, scale-level $d = 0.37$) and absenteeism–lost time criteria (single-item $d = 0.07$, scale-level $d = 0.48$), but effect sizes did not vary markedly for single-item versus scale-level measures of task (ds = 0.19 and

Table 7
Black–White Mean Differences in Job Performance for Scale-Level Data: Criterion Type × Measurement Method × Data Source

Distribution analyzed	d	k	N_{Total}	N_{White}	N_{Black}	90% CI	PVA	$d_{corrected}$
Task (all subjective)	0.25	37	5,882	4,089	1,793	−0.05 to 0.55	43	0.33
Journal	0.25	13	4,428	3,007	1,421	0.05 to 0.45	45	0.32
Technical report	0.25	23	1,392	1,059	333	−0.26 to 0.76	42	0.33
Book	0.40	1	62	23	39			0.52
Contextual (all subjective)	0.14	22	2,093	1,601	492	−0.31 to 0.60	36	0.19
Journal	0.09	9	1,240	890	350	−0.26 to 0.44	39	0.11
Technical report	0.23	13	853	711	142	−0.31 to 0.76	37	0.29
Personality–applied social skills (all subjective)	−0.02	29	4,755	3,336	1,419	−0.55 to 0.52	19	−0.02
Journal	−0.15	3	2,995	1,964	1,031	−0.58 to 0.27	6	−0.20
Technical report	0.22	26	1,760	1,372	388	−0.27 to 0.70	41	0.28
On-the-job training	−0.81	3	479	341	138	−2.41 to 0.78	3	−1.05
Overall job performance (all subjective)	0.37	267	52,202	40,240	11,962	0.04 to 0.71	33	0.48
Dissertation	0.29	5	737	414	323	0.29 to 0.29	100	0.37
Journal	0.26	15	5,672	4,661	1,011	0.00 to 0.52	30	0.33
Conference paper	0.57	3	421	327	94	0.57 to 0.57	100	0.74
Technical report	0.39	244	45,372	34,838	10,534	0.05 to 0.73	34	0.50
GATB studies	0.40	176	33,868	24,877	8,991	0.07 to 0.74	34	0.52
No GATB	0.35	68	11,504	9,961	1,543	0.00 to 0.69	36	0.45
Work sample	0.41	16	5,214	3,398	1,816	0.08 to 0.73	24	0.50
Subjective	0.43	8	2,744	1,921	823	0.10 to 0.75	24	0.55
Dissertation	0.66	1	75	47	28			0.85
Journal	0.33	5	1,692	956	736	−0.02 to 0.68	21	0.43
Technical report	0.58	2	977	918	59	0.58 to 0.58	100	0.75
Objective	0.39	8	2,470	1,477	993	0.06 to 0.71	26	0.43
Dissertation	0.50	1	82	50	32			0.56
Journal	0.33	3	1,593	842	751	0.05 to 0.60	21	0.37
Technical report	0.49	4	795	585	210	0.13 to 0.85	31	0.55
Job knowledge test	0.53	9	2,216	1,360	856	0.33 to 0.74	53	0.60
Attendance–lost time	0.48	2	194	145	49	0.48 to 0.48	100	0.63

Note. All General Aptitude Test Battery (GATB) data were measured at the scale level. CI = confidence interval; PVA = percentage of variance accounted for by sampling error.

0.25, respectively), contextual (ds = 0.12 and 0.14, respectively), and work sample (ds = 0.44 and 0.41, respectively) criteria. Conversely, the direction of racial effects reversed from single-item to scale-level measures of on-the-job training (ds = 0.45 and −0.81, respectively) and personality–applied social skills (ds = 0.14 and −0.02, respectively).

Table 8
Mean Criterion Effect Sizes and Mean Criterion Cognitive Loadings for Job Performance Criteria

Criterion type	Mean d	Mean cognitive loading	k associated with mean d
Task	0.21	.16	93
Contextual	0.13	.10	31
Personality–applied social skills	0.07	.15	60
On-the-job training	0.05	.37	7
Overall job performance	0.35	.19	302
Work sample	0.42	.33	23
Job knowledge test	0.53	.49	9
Absenteeism–lost time	0.09	.06	20
Salary	0.14	.19	5
Promotion	0.18	.15	7
Accidents	−0.06	.11	6

Note. Commendations–reprimands results are not reported because cognitive loading data were unavailable.

Measurement method addresses whether work performance is measured subjectively with ratings of performance or objectively scored using mechanical or quantified techniques. Evidence provided in Table 5 suggests that measurement method has a relatively low impact on mean racial differences in work performance (R = .10). Summary results for this moderator presented in Table 2 support this conclusion because effect sizes are very similar for subjective (d = 0.28) and objective (d = 0.22) measures of performance. In general, there does not appear to be a clear pattern of measurement method results. For some criteria, such as work samples, effect sizes are larger in magnitude for single-item subjective (d = 0.52) versus objective (d = 0.39) measures. This effect did not extend to scale-level work sample measures (subjective d = 0.43, objective d = 0.39). Similarly, single-item task measures show virtually no difference in effect sizes when moving from subjective to objective measurement (ds = 0.18 and 0.20, respectively). Lastly, absenteeism–lost time criteria, measured at the single-item level, exhibit larger magnitude effects for objective (d = 0.11) versus subjective (d = −0.01) measures of performance.

Discussion

Our investigation represents the largest meta-analysis to date of Black–White mean differences in work performance. We assessed mean racial disparities in performance using a greater number of

effect sizes for a number of criterion-type categories. In addition, we analyzed the effects of several previously unstudied moderators. Very similar to Roth et al.'s (2003) recent meta-analytic study, we found that the Black–White mean difference in job performance is just over one fourth of a standard deviation in magnitude favoring Whites ($d = 0.27$; see Table 2). With respect to job performance, and contrary to earlier meta-analyses (J. K. Ford et al., 1986; Kraiger & Ford, 1985), it appears that mean racial differences in work performance have decreased over the years. Possibly, two concerns, one methodological and the other societal or legal, may underlie this trend.

First, Roth et al. (2003) identified range enhancement (i.e., the collection of data from extremely high- and low-performing employees) as a factor that may upwardly bias racial effect size estimates. Failure to consider this source of inflated mean racial differences in performance in past research may explain why effect sizes are smaller in more recent studies. Second, passage of the Civil Rights Act of 1991 outlawed posttest scoring adjustments and quota hiring (without court order), practices previously used in some work contexts to increase minority representation. The use of differential hiring standards across racial groups appears to exaggerate mean racial differences in job performance (e.g., Bartlett et al., 1977; Kraiger, 1981), so merit-based selection is likely to reduce racial effects on job performance (Maxwell & Arvey, 1993; Silva & Jacobs, 1993).

Moderators of Black–White Mean Differences in Job Performance

Of the five moderators examined in our investigation, the most influential for Black–White mean differences in job performance were, in order, criterion type, the cognitive loading of criteria, data source, measurement level, and measurement method. Collectively, the five moderators accounted for 38% of the variance in effect sizes summarizing mean racial differences in performance. Clearly, we failed to account for 62% of effect size variance. Additional research is needed to examine possible sources of this unexplained variance.

In the sections below, we turn to the pivotal question for scholars and practitioners examining mean racial differences in job performance: Why do Black–White mean differences in job performance exist? During our discussion, we highlight study findings that are relevant to the question at hand.

Why Do Black–White Mean Differences in Job Performance Exist?

It should be noted that our presentation here is largely speculative. All of our conclusions should be subjected to additional research scrutiny. Potential explanations for racial effects on job performance include mean racial differences in cognitive ability, rating bias, opportunity bias, and rating purpose, as discussed further below.

Mean racial differences in cognitive ability. Criterion type and the cognitive loading of criteria were the two most potent moderators of racial effects on job performance. As shown in Table 8, the magnitude of criterion-type effect sizes was strongly and positively correlated with cognitive loading. A logical inference from

these findings is that mean racial differences in cognitive ability may underlie these effects (e.g., Roth et al., 2001).

To the extent that mean differences in cognitive ability correspond to racial effects on performance, a potential organizational response is to develop performance measures that capture both job-relevant cognitive and noncognitive criteria. We found that racial effects on performance are clearly smaller for criteria with low cognitive loadings (e.g., contextual performance, personality–applied social skills) than those with high cognitive loadings (e.g., job knowledge tests and work samples). Several authors have been critical of criterion development efforts that focus too narrowly on cognitive and psychomotor aspects of performance, to the exclusion of job-relevant noncognitive criteria (e.g., contextual performance) that contribute to effective job performance (Goldstein, Zedeck, & Goldstein, 2002; Murphy, 1996; Raymark, Schmit, & Guion, 1997). This critique is especially relevant considering the increased emphasis on team-based work and customer service in many jobs in today's economy.

We offer two caveats concerning the expansion of the job performance domain to include less cognitively loaded criteria. First, supervisors may already include such criteria in their ratings of overall job performance, as reported in research showing that contextual performance ratings are correlated with both formal and informal supervisory evaluations of job performance (Borman & Motowidlo, 1997; Van Scotter, Motowidlo, & Cross, 2000; Werner, 1994). Thus, the usefulness of noncognitive criteria in decreasing Black–White mean differences in job performance may be greatest in organizational settings in which supervisors do not consider noncognitive performance dimensions when rating their subordinates.

Second, organizations must ensure that noncognitive performance factors can be defended as job related. Federal guidelines on employment practices and cumulative court cases in the United States determine what are considered acceptable job performance dimensions (Malos, 1998). Contextual performance is often defined as extrarole behavior and thus might be viewed by the courts as inappropriate criteria. Likewise, Malos (1998) stated that legally defensible performance appraisal systems should include behavioral versus trait-based criteria, which preclude personality traits as job performance measures. The use of job analysis to identify job-relevant noncognitive criteria should increase their defensibility as performance measures (e.g., Raymark et al., 1997).

Rating bias. An additional potential cause of Black–White mean differences in job performance is rating bias. Negative stereotypes of and discrimination against Blacks may lead raters to deflate their ratings, resulting in mean performance differences disfavoring them (T. Cox & Nkomo, 1986; Kraiger & Ford, 1985). Alternatively, assigning low ratings to Blacks may arouse legal concerns about claims of employment discrimination. In response, raters might be motivated to inflate Blacks' ratings, resulting in lower racial effects on performance, possibly even favoring Blacks (Kraiger & Ford, 1985; Mobley, 1982).

Although measurement method failed overall to moderate racial effects on job performance, we obtained slight evidence of variability in effect sizes by method. For work samples, this effect may indicate racial bias against Blacks, as effect sizes were larger for subjective than objective criteria measured at the single-item level ($d = 0.52$ vs. 0.39). Though outside of the job performance domain, measurement method comparisons for academy training

criteria also disadvantaged Blacks. Larger racial effects were obtained for subjective than objective indices, summarized overall ($d = 0.51$ vs. 0.41) and measured at the scale level ($d = 0.62$ vs. 0.41). Conversely, an indication of rating bias favoring Blacks is our finding that effect sizes for absenteeism (measured at the single-item level) were smaller for subjective ($d = -0.01$) versus objective ($d = 0.11$) measures. These conclusions must be qualified, because some of these measurement method comparisons were based on few samples. In sum, organizations should provide contexts that reinforce and reward rating accuracy (Murphy & Cleveland, 1995) to reduce the influence of rating bias on Black–White mean differences in job performance.

Opportunity bias and rating purpose.   Finally, opportunity bias and rating purpose are two plausible determinants of Black–White mean differences in job performance. Opportunity bias is evident when members of racial subgroups differ in their opportunity for work success. Rating purpose concerns whether performance ratings are collected for use in making administrative decisions or for research such as criterion-related validity studies (Murphy & Cleveland, 1995). Rater leniency is typically greater for administrative versus research ratings. Our data set did not allow evaluation of these two determinants here, so subsequent research is needed to assess their impact on mean racial disparities in job performance.

Limitations

There are five limitations of our investigation that warrant attention. First, findings for several criterion-type categories were based on very few effect sizes (e.g., salary, promotion, and accidents). A small number of studies increase the possibility that they are randomly unrepresentative of all possible studies (i.e., second-order sampling error). Also, having few studies restricts the number of hierarchical analyses that can be conducted. Specifically, results from a small number of investigations increase the likelihood that untested moderators might influence findings. Therefore, conclusions regarding these criteria should be viewed as tentative, until more data accumulate.

Second, results reported for the cognitive loading moderator were based on a smaller subset of our overall data set. This occurred because many of the studies included here did not use cognitive ability tests (see Terpstra & Rozell, 1997, for possible explanations). Although the correlated vectors analyses were based on 291 samples, some of our results reported within criterion-type categories were drawn from analyses of few effect sizes. In addition, none of the studies assessed included both cognitive ability and personality test data, so we were unable to compare the moderating effects of cognitive loading and personality loading directly within single studies. This is a topic in need of research. However, we are hopeful that investigators improve their reporting of relevant study statistics to aid future meta-analytic research efforts.

A third limitation is that we did not examine the moderating effects of validation study design (i.e., predictive or concurrent). Most of our data came from concurrent criterion-related validity studies. There are few studies of mean racial differences in job performance outside the context of a concurrent study, which prevented us from sorting studies by validation study design. Future meta-analytic investigators should assess the impact of

study design on mean Black–White differences in work performance.

A fourth shortcoming is that we failed to examine Hispanic–White mean differences in work performance, yet Roth et al. (2003) treated this topic. Similar to these authors, we found it difficult to obtain studies with sizable numbers of Hispanic employees, which prevented us from conducting meaningful moderator analyses.

A final limitation is that our corrections for measurement error were approximate. The corrections did not entertain the probability that the reliability of job performance criteria varies across studies. These differences across studies in measurement error may have contributed to unexplained variance in the distribution of mean racial work performance difference effect sizes. Our inability to more accurately estimate the reliabilities of criteria also led us to interpret observed effect sizes rather than corrected effect sizes. Thus, our findings are conservative estimates of the magnitude of mean racial differences in job performance.

Conclusions

Across criterion categories, Black–White mean differences in work performance are evident. For job performance, the differences are slightly more than one fourth of a standard deviation in magnitude, favoring Whites. Criterion type and the cognitive loading criteria were the first and second most potent moderators of effect sizes, respectively. These two moderators are highly related in that the cognitive loading of criterion type varies with mean racial differences in job performance. Racial mean disparities were also moderated by data source, as effect sizes were generally smaller in published versus unpublished sources, and data level, with larger mean differences occurring for scale-level rather than single-item criteria. Given the importance of the topic, we strongly encourage research into the causes of and possible remedies for mean racial differences in performance.

References

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

*Abalos, A., McDaniel, S., & Kisner, R. F. (2000). *Validity of the HPI for selecting bus operators* (Tech. Rep. No. 203). Tulsa, OK: Hogan Assessment Systems.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117,* 288–318.

*Allen, L. (1983). Evaluating firefighters' performance. *Psychological Reports, 53,* 1219–1222.

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

Arthur, W., Jr., Bennett, W., Jr., & Huffcutt, A. I. (2001). *Conducting meta-analysis using SAS.* Mahwah, NJ: Erlbaum.

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.

*Barnett, G., Shin, H. C., & Holland, B. (2000). *Validity of the Hogan Personality Inventory for selecting crewmen* (Tech. Rep. No. 214). Tulsa, OK: Hogan Assessment Systems.

*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 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.

*Bayless, J. A., Lyons, T. J., Park, R. K., & Hayes, T. L. (2002). *The development and validation of a new battery for selecting entry-level detention enforcement officers at the U.S. Immigration and Naturalization Service* (Rep. No. 02–01). Washington, DC: United States Immigration and Naturalization Service.

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.

Blau, G. (1994). Developing and testing a taxonomy of lateness behavior. *Journal of Applied Psychology, 79,* 959–970.

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

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, 52,* 561–589.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco: Jossey-Bass.

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

*Brinkmeyer, K., & Hogan, R. (1996). *Validity of the Hogan Personality Inventory for selecting delivery contractors* (Tech. Rep. No. 93). Tulsa, OK: Hogan Assessment Systems.

*Brinkmeyer, K., & Hogan, R. (1998). *Validity of the Hogan Personality Inventory for selecting customer service representatives* (Tech. Rep. No. 149). Tulsa, OK: Hogan Assessment Systems.

*Brinkmeyer, K., Hogan, R., & Heidelberg, H. (1997). *Preemployment screening preliminary report for pipe manufacturing workers* (Tech. Rep. No. 136). Tulsa, OK: Hogan Assessment Systems.

*Brinkmeyer, K. R., & Hogan, R. (1997). *Validity of the Hogan Personality Inventory for selecting field representatives* (Tech. Rep. No. 107). Tulsa, OK: Hogan Assessment Systems.

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

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

*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.* Princeton, NJ: Educational Test Service.

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

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

Chung-Yan, G. A., & Cronshaw, S. F. (2002). A critical re-examination and analysis of cognitive ability tests using the Thorndike model of fairness. *Journal of Occupational and Organizational Psychology, 75,* 489–509.

Civil Rights Act of 1991, Pub. L. No. 102–166, 105 Stat. 1071.

*Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86,* 410–417.

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

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

*Cox, T., Jr., & Nkomo, S. M. (1986). Differential performance appraisal criteria: A field study of Black and White managers. *Group & Organization Studies, 11,* 101–119.

*[Criterion-related validation data]. (2003a). Confidential unpublished raw data set.

*[Criterion-related validation data]. (2003b). Confidential unpublished raw data set.

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

*Distefano, M. K., Jr., 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., Jr., Pryer, M. W., & Craig, S. H. (1980). Job-relatedness of a posttraining 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.

*Dudley, N. M., McFarland, L. A., Goodman, S. A., Hunt, S. T., & Sydell, E. J. (2002, April). *Social desirability scales: Do race differences exist?* Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada.

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

*Farr, J. L., O'Leary, B. S., Pfeiffer, C. M., Goldstein, I. L., & Bartlett, C. J. (1971). *Ethnic group membership as a moderator in the prediction of job performance: An examination of some less traditional predictors* (Rep. No. 151–277). Silver Spring, MD: American Institutes for Research.

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

Feldman, J. M., & Hilterman, R. J. (1977). Sources of bias in performance evaluation: Two experiments. *International Journal of Intercultural Relations, 1,* 35–57.

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 relationship.* 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 nonminority telephone service representatives. *Journal of Applied Psychology, 56,* 135–139.

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

Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology, 51,* 357–374.

Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). g: Is this your final answer? *Human Performance, 15,* 123–142.

*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.

Greenhaus, J. H., & Parasuraman, S. (1993). Job performance attributions and career advancement prospects: An examination of gender and race effects. *Organizational Behavior and Human Decision Processes, 55,* 273–297.

*Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal, 33,* 64–86.

Hall, F. S., & Hall, D. T. (1976). Effects of job incumbents' sex and race on evaluations of managerial performance. *Academy of Management Journal, 19,* 476–481.

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

*Harville, D. L. (1996). Ability test equity in predicting job performance work samples. *Educational and Psychological Measurement, 56,* 344–348.

Hauenstein, N. M. A., Sinclair, A. L., Robson, V., Quintella, Y., & Donovan, J. J. (2003, April). *Performance dimensionality and the occurrence of ratee race effects.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

*Hogan, J., Hogan, R., & Rybicki, S. (1995). *Validity of the Hogan Personality Inventory and the Inventory of Personal Motives for selecting marketing personnel* (Tech. Rep. No. 77). Tulsa, OK: Hogan Assessment Systems.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88,* 100–112.

*Hogan, J., Holland, B., & Hogan, R. (1998). *Validity of the Hogan Personality Inventory for selecting mechanics* (Tech. Rep. No. 169). Tulsa, OK: Hogan Assessment Systems.

*Hogan, J., & Michel, R. (1996). *Validity of the Hogan Personality Inventory for the selection of cashiers* (Tech. Rep. No. 85). Tulsa, OK: Hogan Assessment Systems.

*Hogan, J., Michel, R., & Hogan, R. (1997). *Validity of personality measures for entry level jobs: Final report* (Tech. Rep. No. 137). Tulsa, OK: Hogan Assessment Systems.

*Hogan, J., & Rybicki, S. (1997). *Validity of correctional officer selection procedures* (Tech. Rep. No. 119). Tulsa, OK: Hogan Assessment Systems.

*Hogan, R., & Heidelberg, H. (1998). *Validity of the Hogan Personality Inventory for selecting dockworkers* (Tech. Rep. No. 130). Tulsa, OK: Hogan Assessment Systems.

*Holland, B., & Hogan, J. (1999a). *Validity of the Hogan Personality Inventory for selecting clerical support aides II and III* (Tech. Rep. No. 167). Tulsa, OK: Hogan Assessment Systems.

*Holland, B., & Hogan, J. (1999b). *Validity of Hogan Personality Inventory for selecting outside sales associates* (Tech. Rep. No. 179). Tulsa, OK: Hogan Assessment Systems.

*Holland, B., & Hogan, J. (1999c). *Validity of the Hogan Personality Inventory for selecting recreational leaders* (Tech. Rep. No. 168). Tulsa, OK: Hogan Assessment Systems.

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

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86,* 897–913.

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

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340–362.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

*Igbaria, M., & Wormley, W. M. (1992, December). Organizational experiences and career success of MIS professionals and managers: An examination of race differences. *MIS Quarterly,* 507–529.

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

*Jacobs, R. R., Conte, J. M., Day, D. V., Silva, J. M., & Harris, R. (1996). Selecting bus drivers: Multiple predictors, multiple perspectives on validity, and multiple estimates of utility. *Human Performance, 9,* 199–217.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

*Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology, 86,* 984–996.

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

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

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

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

Koslowsky, M., Sagie, A., Krausz, M., & Singer, A. D. (1997). Correlates of employee lateness: Some theoretical considerations. *Journal of Applied Psychology, 82,* 79–88.

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

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

*Kriska, S. D. (1984). *Firefighter selection test validation study for the city of Columbus.* Columbus, OH: Civil Service Commission.

*Landy, F. J., & Farr, J. L. (1975). *Police performance appraisal.* University Park, PA: Pennsylvania State University, Department of Psychology.

*Lefkowitz, J. (1972). Differential validity: Ethnic group as a moderator in predicting tenure. *Personnel Psychology, 25,* 223–240.

*Lock, J. (1995). *Using Hogan Personality Inventory for selecting customer & policy service representatives, data entry operators, and document processors* (Tech. Rep. No. 138). Tulsa, OK: Hogan Assessment Systems.

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

*Lyons, T. J., Reilly, S. M., Beatty, G. O., Valdivia, P., & Park, R. K. (2000). *The development and validation of a new battery for selecting entry-level border patrol agents* (Rep. No. 00–4). Washington, DC: United States Immigration and Naturalization Service.

Malos, S. B. (1998). Current legal issues in performance appraisal. In J. W. Smither (Ed.), *Performance appraisal* (pp. 49–94). San Francisco: Jossey-Bass.

Maxwell, S. E., & Arvey, R. D. (1993). The search for predictors with high validity and low adverse impact: Compatible or incompatible goals? *Journal of Applied Psychology, 78,* 433–437.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of performance determinants. *Journal of Applied Psychology, 79,* 493–505.

*McDaniel, S. (1998). *Validity of Hogan Personality Inventory for selecting supervisors* (Tech. Rep. No. 151). Tulsa, OK: Hogan Assessment Systems.

*McDaniel, S. (1999). *Validity of the Hogan Personality Inventory for selecting sheriff's deputies* (Tech. Rep. No. 166). Tulsa, OK: Hogan Assessment Systems.

*McDaniel, S. (2000). [Validity of the Hogan Personality Inventory for field sales, salaried professional, and managerial jobs] (Tech. Rep. No. 219). Unpublished raw data. Tulsa, OK: Hogan Assessment Systems.

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

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

Moore, C. L., MacNaughton, J. F., & Osburn, H. G. (1969). Ethnic differences within an industrial selection battery. *Personnel Psychology, 22,* 473–482.

*Moore, M. H. (1973). *An investigation of the influence of ethnic group membership on job attitudes and the relationship between these attitudes and job performance.* Unpublished doctoral dissertation, University of Houston.

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

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

Murphy, K. R. (1996). Individual differences and behavior in organizations: Much more than g. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 3–30). San Francisco: Jossey-Bass.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

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

Outtz, J. L. (1977). *Racial bias as a contaminant of performance evaluations.* Unpublished doctoral dissertation, University of Maryland.

*Overman, R. W. (1977). *The effects of rater–ratee similarity on rated job performance as a function of the rating instrument.* Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Phillips, C. V. (2004). Publication bias in situ. *BMC Medical Research Methodology, 4,* 20. Retrieved September 12, 2004, from http://www .biomedcentral.com/1471–2288/4/20

*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 Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

*Powell, G. N., & Butterfield, D. A. (1997). Effect of race on promotions to top management in a federal department. *Academy of Management Journal, 40,* 112–128.

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. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241–258.

Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job 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.

Raymark, P. H., Schmit, M. J., & Guion, R. M. (1997). Identifying potentially useful personality constructs for employee selection. *Personnel Psychology, 50,* 723–736.

*[Report of racial mean job performance data]. (2001). Confidential unpublished Tech. Rep.

*Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Qualifying Test in officer training school selection decisions. *Military Psychology, 8,* 95–113.

*Rosenfeld, M., & Thorton, R. F. (1974). *The development and validation of a police selection examination for the city of Philadelphia.* Princeton, NJ: Educational Testing Service.

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

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

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

*Ross, R., Brinkmeyer, K., & Hogan, R. (1998). *Validity of the Hogan Personality Inventory for selecting department managers and assistant managers* (Tech. Rep. No. 143). Tulsa, OK: Hogan Assessment Systems.

*Ross, R., & Hogan, J. (1999). *Validity of the Hogan Personality Inventory for selecting store managers* (Tech. Rep. No. 175). Tulsa, OK: Hogan Assessment Systems.

*Ross, R., Rybicki, S., & Hogan, J. (1997). *Validity of the Hogan Personality Inventory for selecting office clerks and office managers* (Tech. Rep. No. 142). Tulsa, OK: Hogan Assessment Systems.

Roth, P. L., BeVier, 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.

Roth, P. L., Huffcutt, A. I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88,* 694–706.

Rothstein, H. R. (2003). Progress is our most important product: Contributions of validity generalization and meta-analysis to the development and communication of knowledge in I/O psychology. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 115–154). Mahwah, NJ: Erlbaum.

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

*Rybicki, S., Brinkmeyer, K., & Hogan, R. (1997). *Validity of the Hogan Personality Inventory for selecting customer service representatives, drivers, and delivery and installation/service employees* (Tech. Rep. No. 102). Tulsa, OK: Hogan Assessment Systems.

*Rybicki, S., & Hogan, J. (1996). *Validity of the Hogan Personality Inventory and the Motives, Values, Preferences Inventory for selecting small business bankers* (Tech. Rep. No. 101). Tulsa, OK: Hogan Assessment Systems.

*Rybicki, S., & Hogan, J. (1997). *Validity of the Hogan Personality*

Inventory Form-S for selecting correctional deputy sheriffs (Tech. Rep. No. 120). Tulsa, OK: Hogan Assessment Systems.

*Rybicki, S., & Hogan, R. (1996). *Validity of the Hogan Personality Inventory for selecting customer service representatives and assistant branch managers* (Tech. Rep. No. 99). Tulsa, OK: Hogan Assessment Systems.

*Rybicki, S., & Hogan, R. (1997). *Validity of the Hogan Personality Inventory for selecting facility administrators* (Tech. Rep. No. 118). Tulsa, OK: Hogan Assessment Systems.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432–439.

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

Schmidt, F. L., & Le, H. A. (2004). *Hunter and Schmidt meta-analysis programs.* Iowa City, IA: Author.

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.

*Schmitt, N., Hattrup, K., & Landis, R. S. (1993). Item bias indices based on total test score and job performance estimates of ability. *Personnel Psychology, 46,* 593–611.

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.

*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.* Unpublished doctoral dissertation, University of Maryland Baltimore County.

*Shin, H., Van Landuyt, C., & Holland, B. (2001). *Validity of the Hogan Personality Inventory for selecting telephone sales representatives* (Tech. Rep. No. 256). Tulsa, OK: Hogan Assessment Systems.

Silva, J. M., & Jacobs, R. R. (1993). Performance as a function of increased minority hiring. *Journal of Applied Psychology, 78,* 591–601.

*Stafford, A. R. (1983). *The relationship of job performance to personal characteristics of police patrol officers in selected Mississippi police departments.* Unpublished doctoral dissertation, University of Southern Mississippi.

*Stovall, D., Rybicki, S., Hogan, R., & Hauxwell, R. (1997). *Preemployment screening for cashiers* (Tech. Rep. No. 103). Tulsa, OK: Hogan Assessment Systems.

*Strong, M. H., & Najar, M. J. (1999, April). *Situational judgment versus cognitive ability tests: Adverse impact and validity*. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

*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.

Terpstra, D. E., & Rozell, E. J. (1997). Why some potentially effective staffing practices are seldom used. *Public Personnel Management, 26,* 483–495.

Thompson, D. E., & Thompson, T. A. (1985). Task-based performance appraisal for blue-collar jobs: Evaluation of race and sex effects. *Journal of Applied Psychology, 70,* 747–753.

*Thornton, G. C., & Morris, D. M. (2001). The application of assessment center technology to the evaluation of personnel records. *Public Personnel Management, 30,* 55–66.

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.

Tuzinski, K., & Ones, D. S. (1999, April). *Rater–ratee race effects on performance ratings for understudied ethnic groups*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

*U.S. Department of Labor. (n.d.). [General Aptitude Test Battery data file as released to the U.S. Office of Personnel Management]. Unpublished raw data.

*Van Landuyt, C., & Holland, B. (2001). *Validity of the Hogan Personality Inventory for selecting mechanics* (Tech. Rep. No. 241). Tulsa, OK: Hogan Assessment Systems.

*Van Landuyt, C., Philip, T., & Holland, B. (2001). *Validity of the Hogan Personality Inventory for selecting field service technicians and delivery service representatives* (Tech. Rep. No. 247). Tulsa, OK: Hogan Assessment Systems.

*van Rijn, P. V., & Payne, S. S. (1980). *Criterion-related validity research base for the D. C. Firefighter Selection Test* (PR Rep. No. 80–28). Washington, DC: Office of Personnel Management, Personnel Research and Development Center Examination Services Branch.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81,* 525–531.

Van Scotter, J. R., Motowidlo, S. J., & Cross, T. C. (2000). Effects of task performance and contextual performance on systemic rewards. *Journal of Applied Psychology, 85,* 526–535.

Vosburgh, B. V. (1988). *Police personality and performance: A concurrent validity study.* Unpublished doctoral dissertation, California School of Professional Psychology.

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.

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

Werner, J. M. (1994). Dimensions that make a difference: Examining the impact of in-role and extrarole behaviors on supervisory ratings. *Journal of Applied Psychology, 79,* 98–107.

*Wing, H. (1981). Estimation of the adverse impact of a police promotion examination. *Personnel Psychology, 34,* 503–510.

*Wright, P. M., Kacmar, K. M., McMahan, G. C., & Deleeuw, K. (1995). P=f (M X A): Cognitive ability as a moderator of the relationship between personality and job performance. *Journal of Management, 21,* 1129–1139.