A Re-Examination of Black-White Differences in Work Performance:

The Moderating Effects of Criterion Content, Job Complexity,

Level of Measurement, Measurement Method, Source of Data, and Criterion Cognitive Load

Patrick F. McKay
University of Wisconsin-Milwaukee

Michael A. McDaniel
Virginia Commonwealth University

Paper presented at the 18th Annual Conference of the Society for Industrial-Organizational Psychology, Orlando, Florida, April 12, 2003.

The authors are grateful to Kevin Ford, Kurt Kraiger, and Phil Roth for helpful comments concerning development of this study. We thank our many colleagues for graciously providing data included herein.

Correspondence concerning this article may be sent to Patrick F. McKay, School of Business Administration, University of Wisconsin-Milwaukee, 3202 N. Maryland Avenue, Milwaukee, WI 53201. Electronic correspondence may be sent to pmckay@uwm.edu.

Abstract

This study is the largest meta-analysis to date of Black-White racial differences in job performance. Findings indicate that with the exception of turnover, accidents/violations, and police arrest criteria, Whites consistently outperform Blacks on the job. Our results suggest that racial differences vary considerably based on criterion content. Specifically, work sample and job knowledge test performance showed the largest racial differences and supervisor ratings showed the smallest, perhaps suggesting pressure on supervisors to minimize racial differences in ratings. However, findings for job knowledge tests were based upon few effect sizes (k=7), and both job knowledge test and work sample performance data were collected from primarily trainee samples and may not generalize to more experienced job incumbents. The magnitude of racial differences also was moderated substantially by measurement level, job complexity, cognitive load, and data source. The data source analysis provided strong evidence of publication bias in journals indicating suppression of large magnitude effects. It is recommended that organizations minimize racial job performance differences through merit-based hiring, and after hire, through training and development.

A series of large-scale meta-analyses have shown that Whites generally receive higher mean job performance ratings than their Black counterparts (Chung-Yan & Cronshaw, 2002; Ford, Kraiger, & Schechtman, 1986; Hauenstein, Sinclair, Robson, Quintella, & Donovan, 2003; Kraiger & Ford, 1985; Roth, Huffcut, & Bobko, in press). Mean racial performance differences across these studies, measured in standard deviation units, range from a low of .24 (Chung-Yan & Cronshaw, 2002; excluding GATB-based studies from effect size estimates) to a high of .39 (Kraiger & Ford, 1985), suggesting some variability exists in racial effects across studies.

Recent developments in job performance theory, and the lack of or sparse examination of potentially meaningful moderators of racial effects on job performance (i.e., measurement level, criteria content, measurement method, job complexity, cognitive-loading of criteria, data source) warrant re-examination of racial differences in job performance. In the sections to follow, we discuss these moderators in detail and provide justification for their consideration in the current meta-analysis.

Potential Moderators of Racial Differences in Performance Ratings

A goal of the current investigation is to examine a series of moderators expected to affect the magnitude of Black-White mean differences in job performance. The moderators considered are criterion content, measurement method, measurement level, job complexity, the cognitive load of the criteria, and data source.

Criterion Content

Developments in job performance theory of the past 10 years warrant re-examination of racial differences in mean job performance ratings. Current theories of performance emphasize the multidimensional nature of work. Borman and Motowidlo (1993) proposed that job performance be divided into two broad dimensions, namely task and contextual performance. Task performance is defined as "the proficiency with which job incumbents perform activities that are formally recognized as part of their jobs" (Borman & Motowidlo, 1993, pg. 73). Contextual performance represents non-required tasks that are performed to help individuals in an organization (e.g., assisting co-workers with work tasks), or the organization as a whole (e.g., following organizational rules and procedures; Borman & Motowidlo, 1993). Similarly, Campbell, McCloy, Oppler, and Sager (1993) developed an eight-factor model of job performance that includes both task and contextual factors.

Task versus Contextual Job Performance

Personnel theory and research suggest that lower levels of racial differences may occur for contextual versus task performance. First, Borman et al. (1993) theorized that cognitive ability affects one's proficiency in performing task duties, whereas personality should influence a person's inclination to perform contextual work behaviors. Validation studies have shown, in fact, that cognitive ability measures are more correlated with task performance than with contextual performance, while the trend is in the reversed direction for personality measures (Hattrup et al., 1998; Hurtz & Donovan, 2000; Johnson, 2001; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Motowidlo & Van Scotter, 1994; Van Scotter & Motowidlo, 1996).

Because task performance is best predicted by cognitive ability, greater racial differences in ratings should occur for task performance because cognitive ability tests result in substantial adverse impact against Blacks (Chung-Yan & Cronshaw, 2002; Roth, BeVier, Bobko, Switzer, & Tyler, 2001; Schmitt, Clause, Pulakos, 1996). Personality, on the other hand, is predictive of contextual performance and exhibits less adverse impact against Blacks (Bobko, Roth, & Potosky, 1999; Schmitt, Clause, Pulakos, 1996).

Only one meta-analytic study has investigated whether Black-White differences in mean job performance ratings vary based on the task-contextual performance distinction (Hauenstein, Sinclair, Robson, Quintella, & Donovan, 2003). Hauenstein et al (2003) found that supervisory ratings of task performance (d = .366, k = 10, N = 18,481) resulted in lower Black-White mean performance differences than contextual performance ratings (d = .266, k = 10, N = 1634). However, given the limited number of effect sizes included, criterion content as a potential moderator is worthy of further study with a larger number of effects.

Based on theory and findings that suggest that task performance criteria are higher in cognitive load than contextual performance, Black-White differences in job performance are likely to vary by criterion content. Although racial differences due to criterion content differences could occur across categories of criteria, we offer a specific hypothesis concerning task and contextual performance: Criterion content will moderate Black-White mean differences in job performance such that racial differences will be larger for task performance than contextual performance.

In addition to the task-contextual performance distinction, Cascio (1998) further subdivided the performance domain into performance ratings, quality, output/quantity, absenteeism, turnover, and training proficiency. Roth et al. (in press) used Cascio's (1998) taxonomy to code criterion content in their recent meta-analysis. The authors added additional performance categories including safety and organizational citizenship behaviors (OCBs), although a Black-White effect size was not reported for the OCBs. The literature suggests that racial differences in job performance can be expected for some of the additional performance criteria considered here (e.g., training measures), however it is more difficult to make such predictions for other criteria due to the lack of prior research (e.g., absenteeism). We now describe the extant literature concerning racial differences for each type of performance criterion.

<u>Job Knowledge Tests</u>

Job knowledge tests have been used in the industrial-organizational psychology literature as both predictors and criteria (Cascio & Phillips, 1979; Distefano, Pryer, & Craig, 1976, 1980; Rijn & Payne, 1980; Wing, 1981), though we focus on the latter usage in the current study. Also, job knowledge tests are commonly used as criteria in academy training programs (Distefano, Pryer, & Craig, 1976, 1980; Rijn & Payne, 1980).

Hunter (1983) attempted to explain the effect of cognitive ability on job performance, and reported path analytic findings showing that cognitive ability impacts job performance through the acquisition of job knowledge. Because the development of job knowledge is dependent on cognitive ability, and Blacks obtain lower cognitive ability test scores than Whites (Roth et al., 2001; Schmidt & Hunter, 1998), racial effects that occur for job knowledge tests may favor Whites. Schmidt, Clause, and Pulakos (1996) reported an effect size of -.38 (favoring Whites), though job knowledge tests and work samples were collapsed to derive the *d* estimate. However, the Roth et al (in press) meta-analysis estimated race effect sizes on job knowledge and found an objective *d* (assumed here to be reflective of job knowledge test scores) of .67 (k = 2, N = 698). Currently, we re-examine racial effects on job knowledge tests using a larger number of effect sizes estimates (k = 7).

Work Samples

Work sample measures have been used as predictors in validation studies (Gael, Grant, & Ritchie 1975a, 1975b; Harville, 1996). Often, work samples depend on psychomotor abilities, as opposed to cognitive abilities, as evidenced by lower adverse impact for work samples relative to cognitive ability measures as predictors (Motowidlo, Carter, Dunnette, Tippins, Werner, Burnett, & Vaughn, 1992; Schmitt & Mills, 2001; Schmitt, Pulakos, & Clause, 1996). DuBois, Sackett, Zedeck, and DuBois (1993) used a work sample criterion and found that Whites performed slightly higher than Blacks on a grocery clerk work sample (d = .12 for both speed and accuracy criteria, N = 63 and 114, respectively). Conversely, Roth et al. (in press) reported a large mean standardized difference of .52 on work samples favoring Whites (k = 10, N = 3,651). From the presentation above, it is not entirely clear whether work samples result in large Black-White differences in job performance. In this study, we provide analyses to assess the mean racial effect size associated with work samples with a larger number of studies than previously analyzed.

Personnel Records

For many jobs, personnel records criteria are collected including measures of absenteeism/lost time, accidents, salary, shortages/errors, promotions, etc. Ford et al (1986) first meta-analyzed these forms of criteria, however, Roth et al (in press) noted that Ford et al included few studies from which to cleanly parse these criteria into homogenous categories to estimate effect sizes. Roth et al. culled a greater number of studies enabling them to meta-analyze the Black-White effect sizes for personnel records criteria including lost time, turnover, promotion, and safety and found that standardized racial differences on these personnel data range from .19 (k = 11, N = 2,376) for absenteeism to .31 for promotion (k = 7, N = 1,404). For the purpose of comparison, we also examine the racial effects associated with personnel records criteria as assessed by Roth et al.

Training Criteria

In the industrial-organizational literature, training performance has been treated as a criterion in validation efforts (Bayless, Lyons, Park, & Hayes 2002; Blumberg, Farr, Landy, Neidig, Saal, & Whitaker, 1974; Grant & Bray, 1970; Kraiger, 1981; Kriska, 1984; Roberts & Skinner, 1996). From a review of this literature, it appears that training criteria can be sorted into programs housed in training academies (e.g., Kraiger, 1981; Kriska, 1984), government/military training (e.g., Bayless, Lyons, Park, & Hayes 2002; Roberts & Skinner, 1996), and on-the-job training typical of the skill trades (e.g., Blumberg et al., 1974; Tenopyr, 1967), though some programs include both academy and on-the-job elements (Gael & Grant, 1972; Kriska, 1984).

Ackerman's (1988) theory of skill acquisition proposes three stages of learning in training. In the first stage, the declarative stage, training proficiency is highly dependent on cognitive ability because trainees are required to learn knowledge necessary to perform tasks. In the second stage, knowledge compilation, trainees must use perceptual speed to learn production routines that improve the speed and accuracy of performance. During the third phase, the procedural phase, psychomotor ability is needed for trainees to execute actions required to perform learned tasks effectively.

Academy training programs require academic learning of job information in classroom settings, and use scholastic tests of job knowledge and/or training (classroom) grades to gauge trainee proficiency. These programs can be likened to the declarative stage of Ackerman's (1988) model so training proficiency should be based primarily on trainees' cognitive ability. Research supports this reasoning as cognitive ability is significantly related to both training proficiency (Hunter & Hunter, 1984) and academic performance (Brodnick & Ree, 1995). On-the-job training, because it is more performance oriented, can be likened to the knowledge compilation and procedural stages of learning. As such, success in on-the-job training should be less g-loaded.

Because academic training programs are likely to be more g-loaded than on-the-job training, racial effects on academy training proficiency should be larger than those obtained for on-the-job training. Indirectly, Roth et al.'s finding of an effect size of only .12 for on-the-job training (k = 2, N = 132) supports our reasoning, though this finding was based on only two effect sizes.

Measurement Method

Performance data can be measured objectively or subjectively. Objective data are comprised of quantified behavioral or performance measures such as productivity or number of days missed from work. It is commonly assumed that objective performance data are less biased than subjective data that involve ratings of job performance provided by others such as supervisor ratings (Ford et al., 1986).

Performance appraisal research supports the objective-subjective performance distinction as objective and subjective performance measures have been found to correlate at only .389 (Bommer, Johnson, Rich, Podsakoff, & Mackenzie, 1995). In some instances, racial effects have not been shown to vary based on measurement method. Ford et al (1986) found similar effects sizes for objective (d = .209, k = 53, N = 2817) and subjective (d = .204, k = 53, N = 6791) criteria. Roth et al. (in press) found similar results for objective (d = .24, k = 8, N = 2,538), and subjective (d = .20, k = 10, N = 1,811) ratings of performance quality. However, Roth et al. also reported more disparate objective-subjective effect sizes for job knowledge (objective d = .55, k = 10, N = 2,027; subjective d = .15, k = 4, N = 1,231), performance quantity (objective d = .32, k = 3, N = 774; subjective d = .09, k = 5, N = 312), and absenteeism criteria (objective d = .23, k = 8, N = 1,413; subjective d = .13, k = 4, N = 642).

Because previous research has addressed the objective-subjective criterion dichotomy and its impact on Black-White mean job performance differences, we do not hypothesize about this distinction. Instead, we report effect sizes separately for objective and subjective criteria for comparative purposes with Roth et al's recent study.

Level of Measurement

The effect of measurement level on Black-White differences in performance will be considered during our investigation. Broad versus narrow bandwidth of measurement has implications for reliability of measurement. In general, reliability is maximized to the extent that a measurement device provides a maximum amount of information. Scale-level or composite ratings of job performance that are constructed from single-item ratings of multiple performance dimensions provide more performance information than single-item ratings taken in isolation. As a consequence, scale-level ratings will exhibit greater reliability than single item ratings. To the extent that single-item criteria are unreliable, these may underestimate racial effects on performance ratings (Hunter & Schmidt, 1990).

No previous performance appraisal meta-analysis has examined the impact of level of measurement on racial differences in performance. Because composite job performance ratings are likely to be measured with greater reliability than single-item dimension ratings, it is possible that Black-White differences in job performance will be greater for scale-level ratings than single-item ratings.

Job Complexity and Cognitive-Load

Job complexity is also proposed as a moderator of racial differences in job performance. Recently, Roth, Bevier, Bobko, Switzer, and Tyler (2001) theorized that job complexity would moderate Black-White mean differences in cognitive ability, reasoning that applicants self-select into jobs consonant with their ability level (see Wilk & Sackett, 1996 for similar discussion). Roth et al. (2001) found that Black-White differences in cognitive ability (favoring Whites) varied as a function of job complexity, such that ability differences were inversely related to job complexity (effect sizes = .86, .72, and .63 for low, medium, and high complexity jobs, respectively). In addition, Roth et al. (in press) addressed job complexity as a moderator of Black-White mean job performance differences, and found that effect sizes ranged from .27 for low-medium complexity jobs (e.g., truck driver; $d = .27$, $k = 20$, $N = 11{,}916$) to .32 for medium complexity positions (e.g., skilled crafts; $d = .32$, $k = 6$, $N = 11{,}375$). Six or fewer effect sizes were included in two of three job complexity categories; therefore, their results regarding this moderator were interpreted cautiously. We re-assess job complexity as a moderator of racial effects on job performance including a greater number of effect sizes within each job complexity category, utilizing the low, medium, and high complexity sorting scheme based upon McDaniel (1986).

As reported by Roth et al (2001), it is likely that job complexity will moderate racial differences in job performance ratings such that larger differences will occur in low- versus high-complexity positions. Whereas job complexity moderates racial differences in ability and ability-performance relations, complexity should also affect the range of ability present among job incumbents.

The cognitive-load of criteria is related to the job's cognitive complexity. The criteria for highly complex jobs should be positively correlated with cognitive ability while the criteria for low complexity jobs should have a lower correlation with cognitive ability. Despite the positive relation between job complexity and the cognitive load of criteria, we anticipate opposite effects on racial differences. To obtain a cognitively demanding job, one typically needs to complete rigorous formal education. Thus, one needs to pass a variety of cognitive hurdles prior to obtaining a complex job. Consistent with Roth et al (2001), the average cognitive differences between Blacks and Whites was smaller for highly complex jobs than for less complex jobs. Thus, we anticipate smaller magnitude racial differences for highly complex jobs compared to less complex jobs. However, within any given job, some work-related criteria may be more dependent on cognitive ability than other criteria. For example, task criteria may be more dependent on cognitive skills than contextual criteria, which might be more influenced by temperament than cognitive factors. Thus, criteria with high correlations with general mental ability are likely to show larger racial differences in performance than criteria with lower correlations with general mental ability.

Source of Data

A final issue examined is whether racial effects on job performance vary as a function of the source of performance data. Publication bias has been raised as an issue in meta-analytic research (Durlak, 1995). It is argued that meta-analytic results will be distorted to the extent that data not available to the researchers differs in the magnitude of the effect sizes from data that are available to the researcher. Given the sensitive nature of investigations of racial differences in job performance, the likelihood exists that publication bias may operate in such studies. In response to fear of legal sanctions that may result from disclosure of disparate impact in job performance ratings of minorities, data showing large racial differences are very likely to be suppressed. To address this issue, we examine publication source as a potential moderator of racial differences in performance.

Method

Sample of Studies

The authors 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 (ERIC), and Dissertation Abstracts databases for articles and theses/dissertations that may contain useful data. Keyword searches were performed using the keywords <u>job performance</u>, <u>performance ratings</u>, <u>performance evaluation</u>, <u>criteria</u>, <u>validation</u>, and combination of keywords (e.g., <u>racial differences</u> and <u>job performance</u>, <u>performance ratings</u> and <u>race</u>). Second, we consulted previous meta-analyses to identify studies for inclusion (e.g., Chung-Yan & Cronshaw, 2002; Ford et al., 1986; Kraiger & Ford, 1985; Roth et al., in press). Third, performance appraisal and test validation researchers/practitioners were solicited to provide performance data from unpublished technical reports. Lastly, the authors performed manual searches of industrial-organizational psychology and testing-oriented research journals (e.g., Academy of Management Journal, Educational and Psychological Measurement, Human Performance, Human Relations, Journal of Applied Psychology, Journal of Applied Social Psychology, Journal of Vocational Behavior, Personnel Psychology). Using these methods, a total of 146 studies were identified.

Criteria for Inclusion

Several decision rules were used to determine if studies qualified for inclusion in the meta-analysis. First, studies were included if they contained supervisory ratings of job performance (e.g., overall performance and/or individual dimensional ratings) and/or performance data from personnel records (e.g., days absent, disciplinary actions) with means, standard deviations, subgroup N-sizes, or other statistics such as F-tests enabling computation of effect sizes, reported separately for Black and White employees. Several investigations were not included for failure to provide necessary statistical indices (e.g., Moore, MacNaughton, & Osburn, 1969; Thompson & Thompson, 1985; Vosburgh, 1988).

Second, only data from studies that provided incumbent job performance and/or training performance data were included. Studies that used (1) job applicants (e.g., Cascio & Phillips, 1979), (2) 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), (3) performance ratings generated from videotaped performance (Brugnoli, Campion, & Basen, 1979), and (4) upward, peer, or self-appraisals (e.g., Bartlett, Goldstein, Mosier, Hannan, Buxton, Simmons, and Cooper, 1977 Study 2, peer nominations 2 criterion; deJung & Kaplan, 1962; Cox & Krumboltz, 1958; Grant-Vallone, 1998; Schmidt & Johnson, 1973) were not included in the meta-analysis.

Third, studies were selected for our meta-analysis 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, only data derived from independent samples were included in the study. Utilizing this decision rule necessitated the exclusion of several data sets. Data from sample one from Clevenger, Pereira, Wiechmann, Schmitt, and Harvey (2001) was also used in Pulakos and Schmitt (1996). As a result, sample one data from Clevenger et al. (2001) were not included in meta-analytic estimates. Although we excluded duplicative data stemming from publication in multiple sources, we did include multiple effect sizes from the same sample if they could populate different cells in the moderator taxonomy. Thus, a single sample might contribute an effect to a task criterion distribution as well as a contextual criterion distribution. We judge the most meaningful analyses to be those that disaggregate the pool of effect sizes into cells that are largely homogenous with respect to the levels of the various moderators. Fifth, we eliminated criterion data from our analysis if they failed to describe actual job and/or training performance. Thus, days of vacation and education and training score criteria from Neidt (1968) samples one and three were not included. In sum, using the above selection criteria, one hundred studies were judged as acceptable for inclusion in the meta-analysis. Our 100 studies exceed the 36 studies included in the Roth et al. study.

Data Coding

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

<u>Criterion Content</u>. Using definitions of performance measures given in primary studies, the two authors coded criterion content based on the job performance taxonomy developed by Borman and Motowidlo (1993). Additionally, we used the criterion content coding scheme used by the Roth et al. (in press) meta-analysis, based on Cascio's (1998) performance taxonomy, for comparability purposes. Specifically, we coded criterion measures according to the Roth et al. 11-category scheme, and sorted each performance measure into objective and subjective categories. This enabled us to compare our results with those of Roth et al., although the two sets of findings are not directly comparable as the meta-analyses differ in studies included. However, our coding scheme differed from Roth et al's scheme in several ways.

First, we subsumed ratings of quality, quantity, and job knowledge within the "task" criterion type to the extent that performance data characterized core tasks associated with a given job. Second, OCBs were included as measures of contextual performance. Third, we sorted job knowledge tests into a separate criterion category. We believe that our classification here corresponds to Roth et al's job knowledge-objective criterion category. We note that job knowledge tests used in the evaluation of training academy performance are included here.

Fourth, training measures were subdivided into training academy performance and on-the-job training performance. Training programs conducted at training academies may involve greater academic learning requirements than on-the-job training, likely resulting in higher cognitive load for academy training and greater Black-White performance differences (cf., Hunter, 1983). We considered on-the-job training performance to be a sub-category of job performance. Fifth, salary was considered a separate criterion domain from promotion, as salary increases do not require promotion while the obverse is usually true. Sixth, objective indices of shortages/errors (i.e., number of complaints) were treated as a unique category, whereas Roth et al. considered quality of performance and measurement method (i.e., objective vs. subjective) separately.

Finally, arrests were treated as a separate criterion classification due to Roth et al's finding that police arrest data resulted in large effect sizes favoring Blacks due to their assignment to higher crime patrol areas than Whites (cf. Baehr, Saunders, Froemel, & Furcon, 1971). The criterion content categories used for the study are presented in Table 1.

<u>Measurement Method</u>. Performance data were coded as either gathered from objective, quantified measures of performance (e.g., number of absences) or subjective supervisory ratings of performance. Previous research has shown that Black-White mean differences in job

performance may differ for objective versus subjective measures of performance (Ford et al., 1986; Roth et al., in press).

Measurement Level. For each performance measure, data were coded to distinguish single-item dimension ratings from composite, scale-level ratings of multiple dimensions summed or averaged into a total performance score.

Job Complexity. For each sample, job complexity was sorted using the "data" code derived from the 9-digit Dictionary of Occupational Titles code listed for each position. The data code is an established measure of the complexity of occupations (Rivkin & McDaniel, 1990). The data code ranges from 0 (i.e., management of complex data) to 6 (i.e., handling of simple data). Due to inadequate sample sizes across all data codes, we collapsed the codes as follows: 0-2=high complexity, 3-4=medium complexity, and 5-6=low complexity.

Analysis

The analysis was based on a psychometric meta-analysis of standardized mean differences (Hunter & Schmidt, 1990). We conducted the analyses in SAS using code adopted from Arthur, Bennett, and Huffcutt (2001). Effect sizes were individually corrected for measurement error. Roth et al assumed a reliability of .8 for objective criteria and .6 for subjective criteria. We used these values for objective and subjective criteria measured at the scale level. However, item level reliabilities rarely exceed .25 (Hunter & Schmidt, 1990). To estimate the reliability of a single item measure, we made the assumption that our scale level data were each composed of five items. Then, using the scale level reliabilities of .6 (for subjective criteria) or .8 (for objective criteria), we applied the Spearman-Brown formula to estimate the reliability of single item subjective and objective measures. Thus, all single-item subjective measures were assumed to have a reliability of .23, and all single-item objective measures were assumed to have a reliability of .44. We consider the assumption that all scales are made up of 5 items to be a conservative assumption in that many scales have more than five items. Thus, our item level reliabilities are upper bound estimates of the reliability of single items and the measurement-error-corrected effects sizes are likely to be downwardly biased.

Results

Our primary analyses were structured around a four-dimensional data structure: level of measurement, measurement method, data source and criterion content. Table 2 shows the number of effect sizes in each cell of this four dimensional structure. From this four dimensional data structure, we excluded job complexity because it was not coded for some observations due to missing data, and an added dimension would have greatly increased the size of the tables. Therefore, we analyzed job complexity separately from other moderators in the hierarchical meta-analysis. We also excluded the cognitive loading of the criteria from the four-dimensional data structure hierarchical meta-analysis, as it was missing from 74% of the samples. Job complexity is treated in a separate non-hierarchical meta-analysis. The moderating effect of the cognitive loading of the criteria was analyzed using vector correlations (Glass, McGaw & Smith, 1981; Hunter & Schmidt, 1990; Jensen, 1998, Appendix B).

Table 2 shows that content and method are highly confounded (i.e., statistically dependent) in these data. All criteria in the categories of task, contextual, and overall job performance are assessed with a subjective measurement method. Content and measurement level also are confounded. All but one work sample effect size is at the scale level of measurement. All job knowledge effect sizes are at the scale level. All absenteeism, turnover, salary, promotion, and accident/violation criteria effect sizes are at the single rating level of measurement. This confounding will be explicitly considered in the discussion of the racial difference effects sizes.

In taxonomies of performance criteria, one often distinguishes among training performance, turnover, and job performance because the criteria are considered to measure distinct performance constructs. For example, McDaniel, Whetzel, Schmidt and Maurer (1994), in their interview validity meta-analysis, drew this distinction in partitioning their criteria. Results for training criteria are presented in Table 3, turnover in Table 4, and job performance in Tables 5, 6 and 7. Table 8 presents results for the job complexity moderator. In each of these tables, each row of the table details the results of an analysis. The first column describes the distribution of effects analyzed. The mean *d* is given in column two and the number of effects (*k*) in the analyzed distribution is in column 3. Columns four, five, and six present the total sample size, the White sample size, and the Black sample size, respectively. Column seven presents the 90% confidence interval for the observed distribution. Column eight (*PVA*) presents the percent of the observed variance that is expected based on sampling error. The last column ($d_{corrected}$) is the mean of the *d* individually corrected for measurement error in the criteria. For all effect sizes, a positive mean *d* indicates Whites had better performance than Blacks.

Although we are reporting the corrected effect sizes, we will limit our discussion to the observed mean effect sizes as more conservative (downwardly biased) estimates of race disparities in job performance. When considering the observed mean effect sizes, the reader should be informed that these observed effect sizes are underestimates of the actual magnitude of racial differences in job performance.

Table 3 presents the presents results for each of the four dimensions of the data structure for training performance criteria. Table 4 presents results for turnover criteria. Table 5 presents the results for the four data structure moderators separately for job performance criteria. Tables 6 and 7 present results for each data structure dimension crossed with every other dimension producing a hierarchical meta-analysis for all job performance criteria. We began the hierarchy by separating the data by level of measurement. Table 6 presents the results for all effect sizes of criteria consisting of single ratings.

Table 7 presents the same results for criteria measured at the scale level. In each of these tables, we present additional analyses for each content area for which data are available at the table's level of measurement. If a cell contained at least 20 effect sizes, we further divided the cell by measurement method (subjective vs. objective). If these cells contained at least 20 effect sizes, we further divided the cell's effect sizes by data source. An example will help illustrate this decision rule. In Table 7, there were 37 work sample effect sizes based on scale data. Whereas there were at least 20 scale level work sample effect sizes, we divided the 37 effect sizes into objective and subjective categories. There were 30 scale-level, work sample, objectively measured effect sizes, so we further divided them by data source. Whereas there were only seven scale level work sample effect sizes that employed subjective measurement methods, we did not further divide these seven effect sizes into data source categories. This data analysis decision rule restricts most analyses to distributions of reasonable size.

We dropped the two police arrest effect sizes from the analysis. The effect sizes for these data were -.45 and -1.01. These data suggest that Blacks were assigned to high crime neighborhoods that afforded greater opportunity to make arrests. Whereas, the arrest rate differences between Blacks and Whites were most likely due to different patrol areas, rather than racial differences, we did not include these effects sizes in the job performance analyses (Baehr, Saunders, Froemel, & Furcon, 1971). Table 8 examines the job's cognitive complexity as a potential moderator. Table 9 presents an analysis of whether the criteria's g loading affected the magnitude of the effect. Table 9 also expresses the strength of the various moderators examined as a correlation (or multiple correlation). Table 9 also includes a multiple correlation summarizing the strength of all moderators simultaneously in predicting racial differences in job performance. Our findings and interpretations of effect size estimates of racial differences in job performance are presented in the discussion below.

Discussion

The authors undertook this study with the hope that racial effects on job performance are not as large as reported in previous meta-analyses. Given new conceptualizations of job performance, unexamined moderators, and measurement-related issues, our aim was to clarify the extent of Black-White differences in job performance. On the upside, summarizing across all criterion content categories (excluding training data in Table 3), observed effect sizes racial effects on job performance are not very large (d = .21). Similarly, ratings of overall performance, aggregated across single-item and scale-level ratings, resulted in an effect size of .27, though scale-level effects (d = .28) are larger than single-item effects (d = .21). Thus, for commonly used job performance measures, Black-White differences in job performance are smaller than effects reported for cognitive ability tests (Chung-Yan & Cronshaw, 2002). These findings are encouraging because performance data often form the basis of promotion decisions in organizations (Heneman & Judge, 2003). Thus, advancement opportunities for Blacks may not be highly adversely affected by the use performance data to determine job promotions.

However, Black-White differences in job performance were moderated to a large extent by criterion-content, cognitive loading, and job complexity. In addition, effect sizes were large for training, work sample, and job knowledge criteria. There is also evidence of publication bias in the data drawn from journals. In the presentation to follow, we address training and turnover criteria, and then discuss the magnitudes of job performance data effect sizes as a function of the four data dimensions (i.e., level of measurement, measurement method, data source and criterion content). We then examine job complexity, and g-loading of the criteria as moderators of the magnitude of racial job performance differences.

Are there racial differences in training performance?

Yes.

Table 3 presents training performance data. These criteria reflect performance in formal job training programs and do not include criteria collected during on-the-job training programs or job knowledge test measures. Training performance criteria show some of the largest racial effects of any criteria in our study. We speculate that there are two reasons for the large magnitude of the racial differences in training performance. First, training academy programs often emphasize the learning of large amounts of material in a small amount of time. This is likely to make successful training performance more cognitively demanding than most work-related criteria. Whereas there are large differences in cognitive ability on average between Blacks and Whites (Roth, et al., 2001), one might expect large racial differences in cognitively demanding performance criteria, such as training academy performance.

Second, training criteria are primarily available for entry-level public safety jobs in the police and fire services. The government organizations that hire police and fire personnel are sometimes under tremendous pressure to increase minority representation. Selection procedures for these jobs often seek to minimize the cognitive-load of the selection battery to minimize racial differences in hiring. These practices would include banding, using cognitive measures as a pass/fail screen, and excluding cognitive measures from the test battery altogether. Efforts such as these that are successful in minimizing racial differences in selection methods may result in substantial cognitive differences between the hired Blacks and Whites (Silva & Jacobs, 1993). These differences possibly exacerbate racial disparities in training performance. Thus, if cognitive skills are not assessed in the selection process, racial differences in training performance and other cognitively-loaded performance criteria may occur.

We note that objective criteria yield a much larger racial difference (d = .70) than subjective criteria (d = .46). The possibility exists that the human raters who provide subjective ratings are under substantial legal and organizational pressure to minimize racial differences in training performance (Mobley, 1982). Objective criteria, on the other hand, are not under rater control, and as a result, these criteria are less influenced by political concerns that raters may face (Longnecker, Sims, & Gioia, 1987; Murphy & Cleveland, 1995). However, it should be noted that high-complexity jobs were underrepresented for academy training criteria. Training performance racial differences for higher complexity job are expected to be lower as greater pre-selection on cognitive ability is likely in such jobs.

Are there racial differences in turnover?

Yes.

Table 4 presents turnover results. There were 10 turnover effect sizes with a mean of -.22. This indicates that Blacks had lower turnover than Whites, on average. There are several potential explanations for this finding. First, because income differences exist between Blacks and Whites (U.S. Census Bureau, 2001), financial needs related to employment may be greater for Blacks than Whites. As a result, Blacks may incur greater economic losses for quitting than Whites, thus decreasing their propensity to turnover. A second explanation of this finding is that Blacks may perceive reduced employment opportunity when changing jobs due to perceptions of discrimination in hiring (Braddock & McPartland, 1987; Inniss & Feagin, 1989). Thus, turnover among Blacks may be lower because of the perception of fewer job opportunities available resulting from racial discrimination. Third, to the extent that educational and job skill differences exist between races (Hunter, 1986; Schmidt, 1988), Blacks may encounter reduced job mobility relative to Whites that may reduce Blacks' propensity to turnover. The authors disagree on the credibility of the fourth possible explanation for this difference. The explanation for this difference is argued to be unreasonable by the first author and plausible by the second author. Fourth, in many organizations, it is likely to be more difficult to fire a low performing Black employee than a low performing White employee. The act of firing a Black employee, more so than firing a White employee, may be perceived as an act of racial discrimination (Ford et al., 1986; Mobley, 1982). Because the turnover distribution contained only 10 coefficients, conclusions regarding racial differences in turnover should be viewed as tentative pending the accumulation of additional data. It would be most useful if future race and turnover studies distinguished between voluntary and involuntary turnover.

Does level of measurement influence the magnitude of racial differences in job performance?

Yes.

We defined level of measurement as whether the job performance measure was a single rating or a scale. A scale would be a composite of two or more single ratings. Single ratings will typically have much lower reliabilities than scales. Hunter and Schmidt (1990) have noted that single ratings seldom have reliabilities greater than .25. Large sample evidence indicates that job performance scales typically have reliabilities in the range of .50 to .60 (Rothstein, 1990; Viswesvaran, Ones, & Schmidt, 1996).

Our data reveal evidence that job performance measured at the scale level has larger magnitude racial differences than single item scales. From Table 5, the magnitude of racial differences is more than twice as large for scale measurements than single items (.30 vs. .14). The large difference is also clear when comparing the same criterion content across the two measurement levels. For instance, scales show higher magnitude effects than single ratings for task criteria (34 vs. 16), contextual criteria (.16 vs. .09), and overall job performance (.28 vs. .21). These comparisons cover the bulk of the data in our study. Comparisons based on on-the-job training, and work samples do not support this conclusion, but these contrary findings can be questioned due to the small number of coefficients in the comparison distributions (i.e., nine on-the-job scale-level effect sizes, and one item-level work sample effect size). Past studies of racial differences in job performance (Chung-Yan & Cronshaw, 2002; Ford et al., 1986; Kraiger & Ford, 1985; Roth et al., in press) have not explicitly made the distinction between item level and scale level data. In this research area, as in most other areas, important measurement issues are often overlooked (Cone & Foster, 1991).

Does criterion content influence the magnitude of racial differences in job performance?

Yes.

Comparisons of training criteria (Table 3), turnover criteria (Table 4), and job performance criteria (Tables 5, 6 and 7) indicate varying degrees and direction of racial differences. On-the-job training criteria show moderate racial differences favoring Whites.

Turnover criteria show moderate racial differences favoring Blacks. Job performance criteria show varying ranges of racial differences depending on the job performance content category with almost all categories favoring Whites. Whereas most of our data are job performance criteria, we discuss the differences among the performance criteria content categories in more depth.

A distinction between task and contextual performance is evident in Tables 5, 6 and 7. In Table 5, task criteria are shown to have an observed mean *d* of .17 compared to .10 with contextual criteria. The larger racial differences in task versus contextual criteria are seen in all comparisons in Tables 6 and 7. Across both levels of measurement and all data sources, task criteria show larger racial differences than contextual differences. This finding corroborates theory and research findings based on Borman and Motowidlo's (1993) criterion taxonomy (cf., Motowidlo & Van Scotter, 1994; Borman & Motowidlo, 1997). These authors theorized, and past findings suggest, that task criteria are more g-loaded than contextual criteria. Correspondingly, racial effects on performance are greater for task performance than contextual performance. In sum, racial effects observed in overall job performance seem to be primarily the result of disparities in ratings of task performance rather than contextual performance.

Work sample criteria show some of the largest racial differences (.42, Table 5). For criteria measured by scales, work sample criteria show large effects for subjective criteria (.41) and objective criteria (.43) as shown in Table 7. We note with dismay that work sample measures are typically considered the most content-valid measures of job performance in contrast to supervisory ratings, which may be influenced by non-job-related factors such as racial bias (Ford et al., 1986). The fact that work sample measures yield substantially higher racial differences than task criteria (.16, Table 5; .34, Table 7), contextual criteria (.09, Table 6; .16, Table 7), and overall job performance criteria (.21, Table 6; .28, Table 7), suggest that some supervisors may manipulate job performance ratings to minimize race disparities in performance. Such a practice may result from motivation to reduce claims of racial discrimination against organizations (Ford et al., 1986; Mobley, 1982). This is an unpleasant scenario that may explain the observed data.

An alternative explanation of the large race effect on work sample performance is that these data were often collected among employees-in-training or new hires (e.g., Gael, Grant, & Ritchie, 1975b; Kahn, 1977). Most of the data from other criterion categories (excluding training academy, work sample, on-the-job training, and job knowledge test criterion categories) were collected from concurrent validation samples. It may be that the samples of new hires contained greater performance variability than concurrent samples, due to less restriction in range among trainee samples (similar to predictive samples) than experienced incumbents. Thus, work samples effect sizes may be due to greater range of performance among new hires. Research has shown that job experience is positively related to work sample performance (Schmidt, Hunter, & Outerbridge, 1986). This suggests that racial effects on work sample performance may change over time due to decreased variability in performance that may occur with increased work experience. Similarly, simplex patterns of decreasing ability-performance relations suggest that variability in performance should decrease with time due to learning on the job (Ackerman, 1988), a factor that may affect the magnitude of racial effects on work sample performance. Also, work sample data were not available for high-complexity jobs, suggesting that racial effects on work samples may be smaller in contexts where pre-selection on cognitive ability is more likely. Still, there may be other explanations for our findings that we have failed to consider, and we welcome discussion of such explanations.

Job knowledge criteria typically are considered to be highly content-valid measures of employee knowledges and skills. In our data, these measures show large racial differences (.51, Tables 5 and 7), although the magnitude of the effect is best viewed as tentative given the small number of effect sizes (k = 7). Some may consider the large effect sizes for job knowledge test measures as suggestive that supervisor ratings of task, contextual, and overall job performance underestimate racial differences in job performance, due to legal, organizational, and political concerns among raters.

Similar to work samples, however, job knowledge tests were used primarily with trainee samples (cf., Distefano, Craig, & Pryer, 1976, 1980). As such, trainees are likely to exhibit greater range in job knowledge than experienced incumbents, and the bulk of data in other criterion content categories were collected from incumbents with correspondingly smaller race effect sizes. In addition, job knowledge tests were used primarily with medium- to low-complexity jobs; therefore, racial differences in job knowledge test performance may be lower in high-complexity jobs. A final explanation that the first author endorses and second author does not, is that racial differences in job knowledge test scores may partially reflect lower test-taking motivation among Blacks compared to Whites, a finding associated with cognitive ability testing (Arvey, Strickland, Drauden, & Martin, 1990; Chan, Schmitt, DeShon, Clause, & Delbridge, 1997). Cognitive ability tests are correlated with job knowledge test (r = .45; Hunter, 1986); thus, racial differences in test-taking attitudes may also partially account for racial effects on job knowledge test scores.

In contrast to training academy effect sizes reported previously (d = .47), Black-White differences in on-the-job training performance were appreciably smaller (d = .20). It seems that the lower cognitive load (and less academic nature) of on-the-job training measures compared to training academy performance may explain why racial differences in on-the-job training performances are smaller in magnitude than those reported for training academy performance. Another possibility is that training academy performance relies on some objective measures (e.g., classroom grades) but on-the-job training is almost always a supervisor evaluation. Earlier, we speculated that supervisors might be under pressure to minimize racial differences in performance.

With the exception of number produced (or productivity) criteria (d = .36), racial differences in performance for remaining measures of personnel data were small to moderate. Absenteeism/lost time criteria show some of the smallest racial differences in job performance favoring Whites (d = .07, Tables 4 and 6). The results are close for subjective and objective assessment of this criterion (.05 vs. 08). Black-White differences in job performance were also low in magnitude for salary (d = .08), and larger for shortages/errors (d = .12) and promotions (d = .13). A small effect size favoring Blacks was obtained for accidents/violations (d = -.08). However, racial effects reported for number produced, salary, and promotion criteria were based on few effect sizes, and results for these criteria should be viewed with caution.

Does measurement method influence the magnitude of racial differences in job performance?

Maybe.

We define measurement method as a dichotomy: whether the measure was based on a human rater (subjective) or whether the criterion was objectively assessed. Because of the confounding of criterion content with measurement method, there are few comparisons that can be addressed to answer the measurement method question. Table 5 shows that the mean observed effect size was about the same for subjective and objective criteria (.21 vs. .22). However, Table 6 shows that for single item ratings, objective measures of absenteeism and lost time have slightly larger effect sizes than do subjective measures (.08 vs. .05). Also in Table 6, objectively measured work sample performance has a slightly larger effect mean effect size than subjectively measured work sample performance (.43 vs. .41). Roth et al. (in press) argued that objective methods yield higher effect sizes than subjective methods. Although we find a difference in magnitude with objective criteria yielding larger effects than subjective criteria, the magnitude of the differences are not large. Based on our analyses, we conclude tentatively that effect sizes from objective criteria are larger than effect sizes from subjective criteria. This conclusion is tentative because the confounding of method with content makes it difficult to isolate method effects. The difference we attribute to measurement method might be primarily a function of differences in criterion content. This issue cannot be resolved until more data accumulate with less statistical dependence between measurement method and content.

Does data source influence the magnitude of racial differences in job performance?

Yes.

Data source analyses are relevant to the issue of publication bias. Publication bias occurs when the data available to the researcher are not representative of all results (Durlak, 1995). Most discussions of publication bias address issues of the repression of small effect sizes. In many literatures, small effect sizes may be less likely to be published or otherwise available to researchers because the effect sizes are less likely to be statistically significant. In many disciplines, in most journals, findings with statistically significant results are thought to be more likely to be published. However, in studies of racial differences in job performance, small effect sizes are more socially desirable and more legally and politically advantageous than large racial differences. Thus, to the extent that publication bias exists in this literature, it should serve to suppress the availability of studies showing large effect sizes.

We know that we did not receive all available data. A representative of a large city government indicated that she thought this study was very important to conduct and would consider releasing data, but then proved non-responsive to our request for the city's data. An employee of a consulting firm agreed to provide us data and then declined to respond to additional inquiries about the availability of her data. An employee of a utility sought to release his company's data, but the company's lawyer halted the employee's effort to provide us the data. Parties that did release unpublished data to us were often exceedingly cautious about how the data were to be represented. Consulting firms that released data to us insisted that the name of the company in which the data were collected not be revealed. While most consulting firms are concerned about the confidentiality of any type of data from their clients, we perceived particular sensitivities concerning the race data. Thus, it is clear that the release of data showing racial differences in job performance is a very sensitive subject.

Our data clearly show differences in the magnitude of effects across data sources. Our data show that the effects were substantially smaller when reported in journals (.17) or a book (.00), than from a dissertation (.25), a conference paper (.21) or a technical report (.23). Whereas there was only one book that provided data (Kirkpatrick, Ewen, Barrett, & Katzell, 1968), we do not seek to generalize from these results to all books. We are very concerned about the results contrasting the journal data to data obtained from dissertations, conference papers, and technical reports. The lower magnitude of effects from journals is seen for task and contextual performance regardless of the level of measurement. The effect is also seen for overall job performance for single ratings and scale ratings, with one exception. Scale level conference paper effects have an observed mean of .20 while journals have a mean effect of .22. This exception from the strong overall trend is inconsequential as the effect size estimate from the conference paper results is based on only five coefficients.

The low magnitude of effects for journals relative to other research outlets is noteworthy. We believe the most reasonable explanation for this finding is a publication bias cause stemming from authors' reluctance to submit data with large racial effect sizes to journals for publication. However, there are other alternative explanations. The publication bias in journals may be due to journal reviewers and journal editors influencing the authors to remove racial differences from papers when such differences are large, and encouraging the reporting of racial differences when they are small. Although this explanation is possible, we doubt its credibility because it is inconsistent with our personal experience with reviewers and editors. Another possibility is that journals have higher study quality than dissertations, conference papers, and technical reports, and study quality influences the magnitude of the racial differences. We agree that journal articles, on average, may have higher quality than non-journal articles but we fail to see the mechanism by which high study quality would yield smaller effects. The most relevant study quality issue concerns measurement error. That is, some job performance measures have greater reliability than others (Viswesvaran et al., 1996). However, if the reliabilities of job performance measures from journal articles are higher than from other data sources, the magnitude of the racial effects should be higher in journals, not lower. Therefore, we believe that the most reasonable conclusion is that racial difference data from journal sources sharply underestimate the magnitude of racial differences in performance due to publication bias stemming from authors declining to report racial differences when they are large.

Does the cognitive complexity of the job relate to the magnitude of racial differences in job performance?

Yes.

The cognitive complexity of the job refers to the cognitive demands required for successful job performance. Thus, airline pilots can be considered to have greater cognitive complexity than poultry eviscerators. Table 8 shows the effects of job complexity on the magnitude of racial differences. Jobs of high and moderate complexity have lower racial differences in job performance (.21 and .22) then do jobs of low complexity (.30). We speculate that the differences between lower and higher complexity jobs are a function of two factors. The first factor concerns the screening for higher-level jobs. Most high complexity jobs require the possession of college or graduate degrees. Self-selection by individuals into and out of college and minimum standards of educational institutions will cause Black and White college graduates to be more similar in job-related skills, such as cognitive ability and job knowledge, than Blacks and Whites in the general population. This greater similarity of Blacks and Whites in job-related skills should result in smaller racial differences in higher complexity jobs than in lower complexity jobs. This speculation is consistent with the Roth et al (2001) finding that Black and White incumbents in higher complexity jobs exhibit smaller differences in cognitive ability than incumbents in lower complexity jobs. The second factor concerns the consequences of errors. Job performance errors in highly complex jobs may be more serious than job performance errors in low complexity jobs. Thus, an airline pilot who is prone to many job errors is more likely to be terminated or voluntarily removed from a job than a poultry eviscerator who is prone to make many job errors. The greater willingness of employers to remove poor performing employees from higher than lower complexity jobs would serve to reduce racial differences in job performance at higher complexity jobs.

Does the correlation between cognitive ability test scores and the criteria relate to the magnitude of racial differences in job performance?

Yes.

For this analysis, we used a vector correlation approach to meta-analysis (Glass, McGaw & Smith, 1981; Hunter & Schmidt, 1990; Jensen, 1998, Appendix B). In this approach, one correlates two vectors where the unit of analysis is the study. In our analysis, the first vector is the standardized mean differences measuring the magnitude and direction of racial differences in performance. For this analysis, we included all subcategories of job performance, turnover, and training performance criteria. The second vector consists of the correlation between a general cognitive ability measure and the performance criteria. We conceptualize this correlation as the g-load of the criteria. Whereas most of our studies lacked the correlation between a general cognitive ability measure and the criterion, this analysis was restricted to the 272 observations in which such data were available. The sample-size-weighted correlation between the g-loading of the criteria and the racial difference on the criteria was .30. This result indicates that racial differences in performance increase as the cognitive load of the criteria increases.

What is the relative strength of the various moderators examined in influencing the magnitude of racial job performance differences and what is the magnitude of the joint moderating effect?

Table 9 shows the correlation of each of the five moderators of the racial difference effect size. In each analysis, the observations were weighted by sample size. The correlation for measurement level, measurement method, job complexity and criterion g-loading is a Pearson correlation coefficient. Whereas criterion content and data source are nominal level variables, they were identified as classification variables in a general linear model and the correlation is a multiple correlation. The results indicate that criterion content, criterion g-loading, and measurement level were the top three predictors of racial differences in performance. The data source correlation of .11 is primarily driven by the apparent publication bias in the data from journals and the very aberrant effect sizes from the sole book data source. We note measurement method, which distinguishes between subjective and objective measures, was not a strong moderator, a finding consistent with the hierarchical moderator analyses. Collectively, there is a multiple correlation of .71.

The multiple correlation of .71 indicates that we have explained about 50% of the variance across studies in racial effect sizes. Some of the remaining variance is due to random sampling error and some is due to yet to be discovered moderators. The confidence intervals in Tables 3 through 8 are often large suggesting that non-sampling error variance is typically large. The moderators we examined explained much of this variation. We suggest that the remaining variance across studies is due to differences across organizational standards in hiring practices and the availability of effective training and development opportunities. With respect to hiring, those organizations who apply highly valid selection procedures and race-blind hiring practices should have zero to minimal racial differences in performance. Those organizations who do not use highly valid selection procedures or who consider non-merit criteria (e.g., race) in making hiring decisions should have larger racial differences in job performance. With respect to training and development opportunities, those organizations that offer training programs to improve job skills should be able to reduce extant racial performance differences to the extent that the training is effective and the performance area addressed by training is malleable.

Limitations

There are several limitations of our investigation. First, estimates of racial differences for some criterion content categories were based upon very few effect sizes (e.g., job knowledge tests, number produced, promotion, accidents/violations, etc.). As a result, the findings reported should be viewed as tentative, as stronger conclusions about the magnitude of racial differences in performance for these criteria await future investigation. Surprisingly, we found that few studies reported objective performance data relative to subjective ratings, a problem acknowledged by Ford et al. (1986). Second, the results reported for the effect of criterion g-load on racial differences in job performance were based upon a small subset of our overall data set. This limitation was difficult to avoid as few studies reported correlations between general cognitive ability and the criterion. The smaller number of observations for criterion g-load also caused the analysis of the joint effect of all moderators to be based on a small ($N = 272$) subset of the data.

A third limitation is that for data from test validation studies, we did not examine type of study, predictive or concurrent validation, as a potential moderator of racial differences in job performance. By and large, the studies reported represented concurrent validation studies, which did not allow us to sort studies by validation type to examine this moderator. Although we speculated that some large racial effects on performance might be particular to trainee samples (i.e., work sample, job knowledge tests), we did not test this reasoning directly. Future meta-analytic research should examine the effect of incumbent experience levels as a moderator of Black-White differences in job performance. An investigation of this type would directly address the impact that job tenure has on racial differences in performance. In addition, we did not examine White-Hispanic differences in job performance. Roth et al. (in press) treated this topic; however, similar to these authors, we found it difficult to obtain a large number of effect sizes for Hispanic employees. Given the large number of moderators examined here, we could not obtain the number of effect sizes from Hispanic incumbents necessary to conduct meaningful moderator analyses.

A final limitation concerns the extent of publication bias in our data. Although we presented compelling evidence of publication bias in journals, the other data sources might also be affected by publication bias, although to a smaller extent than the journals. Thus, all the mean effect sizes reported in this paper would be underestimates of the racial differences if publication bias occurred in all the data sources.

Conclusion

On average, Black employees exhibit lower work-related performance than White employees. This is apparent for all criteria except turnover, accidents/violations, and number of arrests. The magnitude of racial effects in job performance varied widely across criterion types, and criteria with lower cognitive loadings (e.g., contextual performance) exhibited smaller effect sizes than criteria with seemingly high cognitive loadings (e.g., training academy performance and job knowledge tests). The results showed that two of the most content valid job performance criteria, work sample performance and job knowledge tests, show substantially larger racial differences than those criteria based on supervisory ratings. However, these findings should be tempered by the fact that the majority of work sample and job knowledge test data were collected from trainee samples in medium- to low-complexity jobs. The large observed effects for work samples and job knowledge tests reported here might be smaller in experienced incumbent samples, and among those working in high-complexity jobs. Alternatively, large effects on these work criteria may signal the influence of legal and political concerns on the performance rating process (Longnecker et al., 1987; Murphy & Cleveland, 1995). Raters may endure legal and organizational pressures to downplay racial differences in work outcomes.

In sum, it appears that racial differences in job performance favoring Whites generalize across a number of criterion content areas. However, it is encouraging that observed racial effects on commonly used measures of job performance such as overall job performance, and measures of task and contextual performance are not exorbitantly large. An implication is that the use of these measures to make personnel decisions (i.e., promotions, pay raises, etc.) may not be severely disadvantageous to Blacks. It is disconcerting that Black-White differences in performance do exist, and the question remains as how to reduce or eliminate the disparities reported. A suggestion by Roth et al. in press may be informative here. These authors proposed that racial differences in job performance might be maximized to the extent that hiring standards differ across races, a possibility likely to occur in some job contexts with pressures to increase minority representation. An implication of Roth et al's suggestion is that employers should use valid selection procedures designed to measure job-relevant skills and abilities as a potential way to reduce racial differences in performance (Maxwell & Arvey, 1993; Silva & Jacobs, 1993). Also, it appears that task performance and number produced (i.e., productivity) criteria may be prime targets for training and development interventions designed to impart job-relevant skills as a means of reducing racial disparities in overall job performance.

References

References marked with an * denote studies that were 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, Fort Collins, CO.

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

Arvey, R.D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. Personnel Psychology, 43, 695-716.

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

*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.

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

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.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M., & Mackenzie, S.B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. Personnel Psychology, 48, 587-605.

Borenstein, M. & Rothstein, R. (1999). Comprehensive meta-analysis. Englewood, NJ: Author.

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.

Braddock, J.H., II, & McPartland, J.M. (1987). How minorities continue to be excluded from equal employment opportunities: Research on labor market and institutional barriers. Journal of Social Issues, 43, 5-39.

*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.R., & Hogan, R. (1997). Validity of the Hogan Personality Inventory for selecting field representatives (Tech. Rep. No. 107). 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.

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. (1996). Group differences and personnel selection decisions: Validity, fairness, and affirmative action. Journal of Vocational Behavior, 49, 122-158.

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. (1998). Applied psychology in human resource management (5th ed). Englewood Cliffs, NJ: Prentice-Hall.

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

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

Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity, and test-taking motivation. Journal of Applied Psychology, 82, 300-310.

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.

*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.

Cone, J.D., & Foster, S.L. (1991). Training in measurement: Always the bridesmaid. American Psychologist, 46, 653-654.

*Confidential technical report (2001).

*Confidential technical report (2003a).

*Confidential technical report (2003b).

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.

Cronbach, J. J. (1990). Essentials of psychological testing. New York: Harper Collins.

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-Organizational Psychology.

Durlak, J.A. (1995). Understanding meta-analysis. In L. Grimm & P. Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 319-352). Washington, DC: American Psychological Association.

*Farr, J.L., O'Leary, B.S., & Bartlett, C.J. (1971). Ethnic group membership as a moderator of the prediction of job peformance. 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, CA.

*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 telephone operators. Journal of Applied Psychology, 60, 411-419.

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

Gaillard, K.M. (1998). The effects of race, sex, and socioeconomic status on the performance ratings of job promotion candidates. Unpublished doctoral dissertation, Auburn University, Auburn, AL.

Glass, G.V, McGaw, B, & Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills: Sage.

*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.

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

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

*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.

Hattrup, K., O'Connell, M.S., & Wingate, P.H. (1998). Prediction of multidimensional criteria: Distinguishing task and contextual performance. Human Performance, 11, 305-319.

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-Organizational Psychology, Orlando, FL.

Heneman, H.G., & Judge, T.A. (2003). Staffing organizations (4th ed.). Middleton, WI: Mendota House.

*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., & 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 the Hogan Personality Inventory for selecting recreational leaders (Tech. Rep. No. 168). Tulsa, OK: Hogan Assessment Systems.

*Holland, B., & Hogan, J. (1999c). Validity of Hogan Personality Inventory for selecting outside sales associates (Tech. Rep. No. 179). 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.

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

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

Hurtz, G.M., & Donovan, J.J. (2000). Personality and job performance: The big five revisited. Journal of Applied Psychology, 85, 869-879.

*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.

Inniss, L., & Feagin, J.R. (1989). The Black ‘underclass’ ideology in race relations analysis. Social Justice, 16, 13-35.

*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. West Port, 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, Houston, TX.

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

*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 OH: Personnel Department, State of Ohio.

*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: Department of Psychology, The Pennsylvania State University.

*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.

Longnecker, C.O., Sims, H.P., & Gioia, D.A. (1987). Behind the mask: The politics of employee appraisal. Academy of Management Executive, 1, 183-193.

*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.

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.

McDaniel, M.A. (1986). The evaluation of a causal model of job performance: The interrelationships of general mental ability, job experience, and job performance. Unpublished doctoral dissertation, George Washington University, Washington, DC.

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

*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, Los Angeles, CA.

*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, Houston, TX.

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

Motowidlo, S.J., & Van Scotter, J.R. (1994). Evidence that task performance should be distinguished from contextual performance. Journal of Applied Psychology, 79, 475-480.

*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., & 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. Colorado State University: Colorado Civil Rights Commission.

O'Connell, M.S., McDaniel, M.A., Grubb, W.L. III, Hartman, N.S., & Lawrence, A. (2002, April). Incremental validity of situational judgment tests for task and contextual job performance. Paper presented at the 17th Annual Conference of the Society for Industrial-Organizational Psychology.

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

*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, Knoxville, TN.

*Plamondon, K.E., & Schmitt, N. (2000). 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-Organizational Psychology.

*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. (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.

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

Rivkin, D., & McDaniel, M.A. (1990). The measurement and validation of occupational aptitude requirements. In A. Lancaster (Chair), The enhancement of the Department of Defense Student Testing Program. Symposium presented at the 98th Annual conference of the American Psychological Association, Boston, MA.

*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. (in press). Ethnic group differences in measures of job performance: A new meta-analysis. Journal of Applied Psychology.

Rothstein, H.R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. Journal of Applied Psychology, 75: 322-327.

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 at Philadelphia, PA.

*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. (1988). The problem of group differences in ability test scores in employment selection. Journal of Vocational Behavior, 33, 272-292.

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.

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: John Wiley & Sons.

*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 an 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, Baltimore, MD.

*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, Hattiesburg, MS.

*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.

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

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.G., & 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 (undated). General Aptitude Test Battery data file as released to the U.S. Office of Personnel Management approximately 1985.

*Van Landuyt, C., & Holland, B. (2001). Validity of the Hogn 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.

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

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

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

Wilk, S.L., & Sackett, P.R. (1996). Longitudinal analysis of ability-job complexity fit and job change. Personnel Psychology, 49, 937-967.

*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.

Table 1. Criterion Content Categories.

Work Performance Criteria

Overall Job Performance-Summary ratings of overall effectiveness on the job.

Task Performance-Ratings of proficiency in performing core duties of a position.

Contextual Performance-Ratings of performance of interpersonal or discretionary behaviors beyond assigned work duties.

Work Sample-Scores earned on tests designed to simulate aspects work tasks.

Job Knowledge Test-Tests of training mastery or knowledge. Includes job knowledge test from training academies.

Absenteeism/Lost time-Subjective or objective measures of attendance and tardiness.

Shortages/Errors-Measures of quality of performance such as Service Complaints.

Turnover-Objective or subjective measures of tenure in or dismissal from a postion.

Number Produced-Measures of the quantity of work completed.

Promotion-Objective or subjective measures of promotions or transfers to other positions.

Salary-Objective measures of salary, merit increases, etc.

Accidents/Violations-Objective or subjective measures of the number of accidents, disciplinary infractions, etc.

Arrests-Objective measures of the number of arrests made law enforcement personnel.

On-the-job Training-Measures of training effectiveness collected on-the-job (e.g., ability to learn).

Training Criteria

Academy Training-Measures of training proficiency collect during academy training (e.g., final course grade). Does not include job knowledge tests.

Table 2. Number of Effect Sizes within Four Dimensional Data Structure: Measurement level by Measurement Method by Data Source by Criterion Category.

Measurement level	Measurement method	Data Source	Task	Contextual	On-the-Job Training	Overall	Work Sample	Training Academy	Job Knowledge Test	Absenteeism/ Lost Time	Turnover	Arrests	Shortages/errors	Salary	Number Produced	Promotion	Accidents/Violations	Row Totals
Single Item Rating	Subjective	Dissertation	66	10		7				8						1	1	93
		Journal article	81	18		10		1		1							1	112
		Conference Paper	25	5		4												34
		Technical report	282	58	10	19		4		3								376
		Data published in textbook	2			4		3										9
		Total Single Rating - Subjective	456	91	10	44		8		12						1	2	624
	Objective	Dissertation					1			2			3	1				7
		Journal article						1		11	7	2	14	1	3	2	2	43
		Conference Paper						4										4
		Technical report			1					9			7	3			3	23
		Data published in textbook					1				3			4		1		8
		Total Single Rating - Objective			1			5		22	10	2	24	9	3	3	5	85
Scale Level Rating	Subjective	Dissertation	7			101												108
		Journal article	23	10		20	4	1										58
		Conference Paper	1			5		3										9
		Technical report	2		6	77	3	19										107
		Data published in textbook	2					10										12
		Total Scale Rating - Subjective	35	10	6	203	7	33										294
	Objective	Dissertation			1		10											11
		Journal article			1		15	1	3				1		1			21
		Technical report			1		5	1	4									14
		Total Scale Rating - Objective			3		30	2	7						1			46
Column Totals		Total Subjective	491	101	16	247	7	41	0	12	0	0	0	0	0	1	2	918
		Total Objective	0	0	4	0	31	9	7	22	10	2	25	9	4	3	5	131
		Total Single Rating	456	91	11	44	1	13	0	34	10	2	25	9	4	4	7	711
		Total scale rating	35	10	9	203	37	37	7	0	0	0	0	0	0	0	0	338
		Total	491	101	20	247	38	50	7	34	10	2	25	9	4	4	7	1049

Table 3. Black-White Differences in Training Performance by Level of Measurement and Measurement Method.

	Criterion	d	k	N_{Total}	N_{White}	N_{Black}	*90% C.I.*	*PVA*	$d_{corrected}$
	All training criteria	.47	50	35,955	32,234	3,721	.08 to .87	9	.75
	Level of Measurement								
	Single item ratings	.47	13	14,725	13,444	1,281	.28 to .65	23	.96
	Scale ratings	.48	37	21,230	18,790	2,440	.00 to .97	8	.61
	Measurement Method								
	Subjective	.46	41	33,313	30,470	2,843	.10 to .81	10	.75
	Objective	.70	9	2,642	1764	878	.12 to 1.29	10	.83
	Data Source								
	Journal	.42	3	26,358	25,125	1233	.30 to .53	9	.73
	Conference paper	.45	7	1,032	526	506	.33 to .57	84	.62
	Technical report	.73	27	7,605	5,917	1,688	.14 to 1.33	11	.94
	Book	.08	13	960	666	294	-.59 to .76	25	.05
	Data Source By Measurement Method By Level of Measurement								
Single item Rating	Subjective	.48	8	13,665	12,912	753	.30 to .66	17	1.00
Single item Rating	Objective	.31	5	1,060	532	528	.31 to .31	100	.47
Scale Level Rating	Subjective	.44	33	19,648	17,558	2,090	.01 to .88	9	.57
Scale Level Rating	Journal	.35	1	12,964	12,453	511	-	-	.45
Scale Level Rating	Conference paper	.63	3	402	213	189	.63 to .63	100	.81
Scale Level Rating	Technical report	.68	19	5,572	4,407	1,165	.07 to 1.30	10	.88
Scale Level Rating	Objective	.97	4	1582	1232	350	.61 to 1.32	19	1.08

Table 4. Black-White Differences in Turnover. All data are Single-Item Ratings and Objective Measurement Method.

Distribution analyzed	*d*	*k*	N_{Total}	N_{White}	N_{Black}	*90% C.I.*	*PVA*	$d_{corrected}$
Turnover (all objective)	-.22	10	2,085	1,367	718	-.74 to .30	16	-.33

Table 5. Black-White Differences in Job Performance by Level of Measurement, Measurement Method, and Data Source.

Distribution analyzed	*d*	*k*	N_{Total}	N_{White}	N_{Black}	*90% C.I.*	*PVA*	$d_{corrected}$
All ratings	.21	987	207,701	151,321	56,380	-.20 to .63	23	.33
Level of Measurement								
Single item ratings	.14	684	112,125	75,652	36,473	-.28 to .55	55	.28
Scale ratings	.30	303	95,576	75,669	19,907	-.07 to .67	20	.38
Measurement Method								
Subjective	.21	877	179,961	133,315	46,646	-14 to .58	29	.34
Objective	.22	110	27,740	18,006	9,734	-.46 to .89	9	.26
Data Source								
Dissertation	.25	219	42,315	25,757	16,558	-.11 to .62	30	.35
Journal	.17	222	57,389	39,295	18,094	-.31 to .65	16	.24
Conference paper	.21	40	34,638	30,887	3,751	.09 to .33	46	.31
Technical report	.23	493	71,186	53,785	17,401	-.24 to .70	26	.40
Book	.00	13	2,173	1,597	576	-.23 to .24	54	.00
Criterion Content Category								
Task	.17	491	77,165	52,741	24,424	-20 to .54	34	.33
Contextual	.10	101	14,709	9,487	5,222	-27 to .46	36	.19
On-the-job training	.20	20	3,212	2,318	894	-80 to 1.20	6	.42
Overall job performance	.27	247	79,544	65,712	13,832	-04 to .58	26	.37
Work sample	.42	38	11,801	7,083	4,718	.07 to .78	22	.49
Job knowledge test	.51	7	1,758	1,104	654	.28 to .75	45	58
Absenteeism/lost time	.07	34	5,939	3,571	2,368	-54 to .68	14	.11
Shortages/errors	.12	25	7,570	5,277	2,293	-.53 to .76	8	.16
Salary	.08	9	1,738	1,287	451	-.29 to .45	29	.12
Number produced	.36	4	940	741	199	-.27 to .99	11	.38
Promotion	.13	4	764	575	189	-.04 to .30	67	.21
Accidents/violations	-.08	7	2,561	1,425	1,136	-.43 to .27	20	-.12

Table 6. Race Differences in Job Performance for Single Item data. Content by Measurement Method by Source.

Distribution analyzed			d	k	N_{Total}	N_{White}	N_{Black}	95% C.I.	PVA	$d_{corrected}$
Task (all subjective)			.16	456	71,431	48,893	22,538	-.21 to .52	34	.33
		Dissertation	.08	66	11,985	6,177	5,808	-.05 to .20	80	.16
		Journal	.06	81	19,341	14,367	4,974	-.34 to .47	22	.13
		Conference paper	.30	25	3,852	2,093	1,759	.08 to .52	59	.62
		Technical report	.22	282	36,113	26,178	9,935	-.14 to .58	40	.46
		Book	.11	2	140	78	62	.11 to .11	100	.24
Contextual (all subjective)			.09	91	12,647	8,148	4,499	-28 to .46	36	.19
		Dissertation	.08	10	1,918	1,164	754	.08 to .08	100	.16
		Journal	.03	18	4,371	2,430	1,941	-.14 to .20	62	.06
		Conference paper	.21	5	880	516	364	.21 to .21	100	.44
		Technical report	.12	58	5,478	4,038	1,440	-.42 to .66	29	.25
On-the-Job Training			.31	11	1,953	1,427	526	-.06 to .68	31	.63
Overall Job Performance (all subjective)			.21	44	7,714	5,286	2,428	-.18 to .59	30	.43
		Dissertation	.29	7	1,203	807	396	.05 to .54	52	.61
		Journal	.23	10	1,791	1,214	577	-.11 to .56	35	.47
		Conference paper	.30	4	591	334	257	.11 to .48	69	.62
		Technical report	.21	19	3,316	2,320	996	-.19 to .61	28	.44
		Book	-.04	4	813	611	202	-.48 to .40	22	-.08
Work Sample (all objective, all dissertation)			.61	1	64	41	23	-	-	.91
Absenteeism/Lost Time			.07	34	5,939	3,571	2,368	-.54 to .68	14	.11
	Subjective		.05	12	2,413	1,244	1,169	-.35 to .44	26	.10
	Objective		.08	22	3,526	2,327	1,199	-.64 to .81	12	.13
		Dissertation	.09	2	218	137	81	.09 to .09	100	.13
		Journal	.16	11	2,476	1,616	860	-.11 to .43	40	.24
		Technical report	-.14	9	832	574	258	-1.49 to –1.21	6	-.21

Table 6 (continued).

Distribution analyzed	d	k	N_{Total}	N_{White}	N_{Black}	95% C.I.	PVA	$d_{corrected}$
Shortages/Errors (all objective)	.08	24	6,904	4,725	2,179	-.56 to .72	8	.12
Dissertation	.08	3	1,661	648	1,003	.00 to .15	77	.12
Journal	.35	14	2,698	1,934	764	-.19 to .89	17	.53
Technical report	-.20	7	2,545	2,143	402	-.82 to .41	7	-.31
Salary (all objective)	.08	9	1,738	1,287	451	-.29 to .45	29	.12
Number produced (all objective)	-.09	3	410	274	136	-.09 to -.09	100	-.13
Promotion (all objective)	.13	4	764	575	189	-.04 to .30	67	.21
Accidents/Violations	-.08	7	2,561	1,425	1136	-.43 to .27	20	-.12

Table 7. Race Differences in Job Performance for Scale Level Data. Content by Measurement Method by Source.

Distribution analyzed	*d*	*k*	N_{Total}	N_{White}	N_{Black}	*95% C.I.*	*PVA*	$d_{corrected}$
Task (all subjective)	.34	35	5,734	3,848	1,886	.01 to .67	39	.44
Dissertation	.66	7	526	328	198	.66 to .66	100	.85
Journal	.27	23	4,391	3,044	1,347	-.04 to .58	37	.35
Conference paper	.61	1	134	71	63	-	-	.79
Contextual (All Subjective, All Journal)	.16	10	2,062	1,339	723	-.14 to .46	37	.21
On-the-job Training	.03	9	1,259	891	368	-1.46 to 1.52	3	.10
Overall Job Performance (all subjective)	.28	203	71,830	60,426	11,404	-.02 to .58	26	.36
Dissertation	.38	101	21,818	14,862	6,956	.06 to .70	33	.49
Journal	.22	20	7,064	5,791	1,273	-.09 to .53	24	.28
Conference paper	.20	5	29,181	27,873	1,308	.12 to .27	27	.25
Technical report	.33	77	13,767	11,900	1,867	-.05 to .70	31	.42
Work Sample	.42	37	11,737	7,042	4,695	.06 to .78	21	.49
Subjective	.41	7	2,597	1,806	791	.12 to .70	26	.53
Objective	.43	30	9,140	5,236	3,904	.05 to .80	20	.48
Dissertation	.50	10	819	499	320	.50 to .50	100	.56
Journal	.43	15	6,747	3,636	3,111	.02 to .84	13	.48
Technical report	.36	5	1,574	1,101	473	.05 to .67	27	.41
Job Knowledge Test	.51	7	1758	1,104	654	.28 to .75	45	.58
Shortages/Errors	.50	1	666	552	114	-	-	.56
Number produced	.70	1	530	467	63	-	-	.78

Table 8. Race Differences in Job Performance by Job Cognitive Complexity.

Distribution analyzed	*d*	*k*	N_{Total}	N_{White}	N_{Black}	*95% C.I.*	*PVA*	$d_{corrected}$
High complexity	.21	384	92,913	74,390	18,523	-.16 to .57	25	.33
Moderate complexity	.22	392	77,502	51,720	25,782	-.22 to .65	23	.32
Low complexity	.30	156	24,512	15,904	8,608	-.11 to .71	30	.43

Table 9. Strength of Each Moderator and the Six Moderator Set in Predicting Racial Differences in Job Performance.

Moderator	R	Sample size
Data level	.27	1047
Measurement method	.03	1047
Criterion type	.44	1047
Data source	.11	1047
Job complexity	.09	983
Criterion g-loading	.30	272
The 6 moderator set.	.71	272