Journal of Management
Vol. 38 No. 2, March 2012 719-739
DOI: 10.1177/0149206310374774

A Meta-Analysis of Gender Group Differences for Measures of Job Performance in Field Studies

Philip L. Roth
Clemson University
Kristen L. Purvis
Cornell University
Philip Bobko
Gettysburg College

There are multiple views in human resource management and organizational behavior concerning gender differences in measures of job performance. Some researchers suggest that males generally are evaluated higher than females across a variety of situations that include job performance measurement. At the same time, the presence of specific status cues in expectation states theory (EST; similar to the concept of individuating information) suggests that measures of job performance will be more similar than different for males and females. Previous analyses are unclear in their results for the measurement of the construct of job performance because they have included, and/or focused on, additional constructs (e.g., hiring suitability, leadership performance aggregated with leadership satisfaction) or have used student samples in lab experiments. The authors of this article conducted a meta-analysis of job performance measures from field studies. They found that females generally scored slightly higher than males (mean d = *–.11, 80% credibility interval of –.33 to .12). Other analyses suggested that, although job performance ratings favored females, ratings of promotion potential were higher for males. Thus, ratings of promotability may deserve further attention as a potential source of differential promotion rates. These findings and processes are discussed within the context of EST.*

Keywords: *gender; job performance; meta-analysis*

Corresponding author: Philip L. Roth, Department of Management, Clemson University, Clemson, SC 29634-1305, USA

Email: rothp@clemson.edu

Introduction

Job performance is an important variable in the human resource management and organizational behavior literatures (e.g., Viswesvaran, Schmidt, & Ones, 2005). One important issue in measuring job performance is whether or not there are stable, persistent differences between ethnic and gender subgroups. While there are two recent meta-analyses relating ethnicity to job performance (McKay & McDaniel, 2006; Roth, Huffcutt, & Bobko, 2003), there has been no recent meta-analytic attention to gender differences in job performance (McKay, 2009). This is unfortunate, as there are multiple views on this issue. One such view is noted in an often-cited narrative review article by Nieva and Gutek (1980: 273). These authors suggested that there is a "fairly consistent" tendency for males to receive more favorable performance evaluations than females receive (see also Greenhaus & Parasuraman, 1993; Igbaria & Baroudi, 1995; Lyness & Heilman, 2006). In turn, higher evaluations of job performance for males could have serious consequences in promotion decisions (Greenhaus & Parasuraman, 1993; Mobley, 1982) and might be a precursor of the glass-ceiling effect (Heilman, 2001; L. M. Shore & Thornton, 1986; T. H. Shore, 1992).

At the same time, there can be a great deal of performance-related information (i.e., performance cues) available to individuals who assess job performance. Expectation states theory (EST) suggests that these specific cues are likely to have a much greater influence on measures of job performance (e.g., ratings) than do cues such as gender. Similarly, there are many psychological similarities between genders, as well as data supporting that view (e.g., Hyde, 2005). Thus, an alternative view is that assessments of male and female job performance may be more similar than different.

Previous meta-analyses have begun to address gender differences on variables related to job performance, but not the measurement of job performance per se. Some meta-analyses have focused on experimental studies of hiring recommendations (Davison & Burke, 2000; Olian, Schwab, & Haberfield, 1988). Others have cumulated mixes of laboratory and field studies while aggregating heterogeneous outcome variables such as leadership effectiveness, satisfaction with leaders, and perceived leadership ability (Eagly, Karau, & Makhijani, 1995; Eagly, Makhijani, & Klonsky, 1992). The different methods, samples (e.g., students vs. incumbents), dependent variables, and mixed results make it difficult to determine if, overall, males and females differ on measures of job performance in operational field settings.

The purpose of this article is to meta-analyze male–female differences for measures of job performance in field settings (e.g., supervisory ratings or measures of output).[1] We do not address differences in salaries, an outcome of performance that has already been meta-analyzed (Ng, Eby, Sorensen, & Feldman, 2005). Instead, we focus on determining if available evidence supports gender similarity in actual or rated levels of job performance in field settings.

We examine "direct" measures of job performance (Cascio, 1997). Such measures include supervisory ratings, measures of quantity and quality, and measures of output, rather than "indirect measures" such as withdrawal behaviors (e.g., absenteeism). We also added measures of promotion potential, because several primary studies contrasted this measure to measures of job performance.

The Importance of Gender Differences in Job Performance

Differences between males and females, on a variety of variables including the measurement of job performance, have received much attention in both the popular press (e.g., MacGillivray, Beecher, & Golden, 2009) and the scientific literature (e.g., Hyde, 2005; King, Hebl, George, & Matusik, 2010). The topic is important for several reasons. In addition to the reasons already noted, the actual relationship between gender and measures of job performance is important for evaluating theories. For example, a widely cited narrative review suggests widespread differences in ratings across gender, such that females are generally disadvantaged (Nieva & Gutek, 1980). Others have suggested that gender incongruity with the perceived role of managers could lead to lower ratings for female managers (e.g., Eagly & Karau, 2002; Schein, 2001) or that females are not perceived as positively as males are in terms of promotability (Lyness & Heilman, 2006).

Gender differences in measures of job performance can influence many other variables within organizational settings. Measured job performance differences could influence both pay and promotions (Robertson, 1986). Or, unjustified differences in performance measures could erode trust (Schoorman, Mayer, & Davis, 2007) or lead to lower levels of job satisfaction (as per Colquitt, Conlon, Wesson, Porter, & Ng, 2001; see also King et al., 2010), which, in turn, lead to withdrawal such as lateness, absenteeism, and turnover (Harrison, Newman, & Roth, 2006). Accurate assessment of the magnitude of any gender differences in the measurement of job performance is also important for the application of human resource models of selection test fairness (Cleary, 1968, see also Bartlett, Bobko, Mosier, & Hannan, 1978). Further, managers deal with a wide variety of demands each day (e.g., Mintzberg, 2001) and may benefit from guidance on where to focus their efforts at creating a fair and balanced workplace.

Previous Meta-Analyses From Related Literatures

Several different disciplines contain literature that contributes to our understanding of, and expectations for, gender group differences in measures of job performance. We briefly review key meta-analyses from the various disciplines.

Experimental studies in personnel selection. Some studies (primarily in applied psychology and judgment and decision making) cumulate experimental studies of hiring decisions. Olian et al.'s (1988) meta-analysis focused on laboratory studies of individuals making decisions or evaluations of hypothetical candidates (see also Davison & Burke, 2000). Olian et al. reported a mean d of .41 ($K = 19$, $N = 1{,}842$), such that males were evaluated, on average, higher than females for hiring-related ratings. Olian et al. also examined the effect of qualifications on hiring-related judgments and found that hiring qualifications accounted for approximately eight times as much variance in the evaluations as did their gender manipulations (more on this below, as we link this analysis to the concept of "individuating information").

Leadership. Two meta-analyses in leadership focused on gender differences. The first meta-analysis focused primarily on laboratory studies (mainly using written or video vignettes

of leaders; see Eagly et al., 1992, including their analysis of the article by Swim, Borgida, Maruyama, & Myers, 1989). The second leadership meta-analysis cumulated lab and field studies (Eagly et al., 1995). The primary studies in this analysis used objective measures of leader performance, including performance in laboratory studies, measures of job knowledge (e.g., tacit knowledge of business), and absenteeism. Subjective measures included self, subordinate, and peer ratings to assess leader effectiveness, leader motivation, leader ability, and satisfaction with the leader (Eagly et al., 1995), although self-report measures by leaders and measures using subordinate reports were most prevalent (see their Table 3). Overall results showed a d of –.02 (K = 76).

Applied psychology and education. Bowen, Swim, and Jacobs (2000) cumulated measures of male–female differences using studies in applied psychology and education that reported information on covariates of the gender–performance relationship (e.g., organizational level, experience, education). The overall mean d was –.01 (K = 32, N not reported). Another set of analyses isolated studies in which there were no self-report measures (mean d = –.05, K = 22, N not reported). The authors interpreted their results as suggesting little overall bias for gender in performance appraisals.

In applied psychology, a large-scale empirical study (Sackett, DuBois, & Noe, 1991) focused on personnel selection within a single, multiple-job, nonstudent database used to evaluate the validity of the General Aptitude Test Battery (GATB). The researchers created 486 firm–job combinations in which they examined male–female differences based on supervisory performance appraisals, and they found that males were rated higher than females (d = .07).

Limitations of previous literatures. Overall, there are a variety of reasons why it is difficult to ascertain an accurate level of gender differences in measures of job performance in organizational settings. First, many results may be limited by a lack of individuating information and a lack of generalizability to actual measures of job performance in organizational settings (Davison & Burke, 2000). This concern applies most clearly to the experimental studies using student participants in personnel selection but is also relevant to the leadership studies (as student samples in laboratory studies were present in both meta-analyses).

Second, some meta-analyses, particularly in the leadership literature, contain a highly heterogeneous set of dependent constructs or variables that includes ratings of leader performance, leader satisfaction, job knowledge, absenteeism, and so on. Thus, it is difficult to isolate the gender–job performance relationship. Relatedly, many jobs may not involve leadership-related tasks and duties, so generalization of this literature to those other jobs is problematic.

Third, several sources of information in the prior analyses are of concern. For example, some ratings in the above analyses are self-ratings, in spite of generalized concern about such ratings (Farh & Dobbins, 1989; Harrison & Shaffer, 1994). Again, this concern is most focused on leadership studies. (In this regard, Bowen et al., 2000, note that more than 90% of performance appraisals in organizations are conducted by supervisors.) Further, in the education literature, we are concerned by use of student ratings of teachers, because such ratings are not generally accepted measures of job performance in human resource management and organizational behavior.

Finally, the fact that managerial and professional jobs were "virtually absent from the database" (Sackett et al., 1991, p. 264) from the GATB study raises concerns. Results based upon the inclusion of many "blue-collar" jobs (e.g., machinist) could readily be confounded by association of the jobs within that database with male stereotypes (see Davison & Burke, 2000).

Theoretical Perspectives and Research Hypotheses

Expectation states theory. EST provides a theoretical framework to understand gender differences in the measurement of job performance (see Berger, Rosenholtz, & Zelditch, 1980; Berger, Wagner, & Zelditch, 1985; Correll & Ridgeway, 2003; Dovidio, Ellyson, & Keating, 1988). EST suggests that gender is one of several "diffuse" (or indirect) status cues that influence expectations about the knowledge, ability, or influence of a given person. Diffuse cues (e.g., gender or ethnicity) can be conceptually linked to performance on many tasks, although the links may not be as strong as with specific status cues. The diffuse cue of gender is generally thought to suggest that females are more artistic and literary, kinder, and more patient and understanding, while men are thought to be more scientific, mechanical, and assertive (e.g., Berger et al., 1980; see also Eagly & Karau, 2002). This can translate into females initially having less status in many work situations (Berger et al., 1985). For example, all else equal, a female in an auto parts store may be judged to have a lower level of job knowledge than a male does.

Specific status cues are typically more directly related to performance expectations than diffuse status cues are. That is, specific cues typically have a fairly short logical chain relating them to expectations (e.g., the cue of spatial ability is directly linked to job performance in air traffic control), whereas diffuse expectations have longer logical chains (e.g., the cue of male might relate to spatial ability, which relates to job performance in air traffic control). When both direct and diffuse status cues are present, EST researchers note that the task-relevant, specific cues usually substantially outweigh the diffuse cues (Dovidio et al., 1988; Freese & Cohen, 1973; W. Wood & Karten, 1986). To continue the above example, the female in the auto parts store may be judged to have superior job knowledge when one learns she has a master's degree in mechanical engineering and is a certified mechanic.

Similarly, social psychologists and sociologists have used the construct of individuating information to understand judgments about gender differences in the workplace (e.g., Reskin, 2000; see also Eagly et al., 1992, 1995, or Olian et al., 1988). Observed performance on the job in question (and/or job-related knowledge, skills, and abilities) could be a particularly powerful source of individuating information for someone assessing job performance. Reskin notes that individuating information can counter stereotypes, given the relevance of such information to many judgmental tasks (e.g., hiring decisions). This use of individuating information is quite similar to the role of specific cues in EST in that individuating information should be more closely related to performance than are diffuse cues. That is, individuating information may minimize the role of dynamics such as stereotypes.

The application of EST, and the related concept of individuating information, has several implications for *field* studies of job performance. In particular, the availability of substantial

amounts of actual performance information can be seen as a potentially large volume of important specific status cues. Thus, in field studies of job performance, one would posit smaller gender differences in measures of job performance (and differences might slightly favor females, given the material noted below). In contrast, in experimental studies on performance evaluation or hiring potential, specific cues (and their variability) are often controlled by keeping them constant. Thus, EST helps predict a difference in results between field studies of job performance assessment and experimental studies.

In sum, given the above logic using EST and the convergence with the concept of individuating information, we hypothesize:

Hypothesis 1: Gender differences on measures of job performance in field studies will tend to be small (e.g., $d = .2$ or less).

Theories of job performance. The theory of job performance by Campbell and colleagues (e.g., Campbell, Gasser, & Oswald, 1996; Campbell, McCloy, Oppler, & Sager, 1993; McCloy, Campbell, & Cudeck, 1994) suggests that three constructs are the immediate precursors of underlying or actual job performance—declarative knowledge, procedural knowledge, and motivation (see also the work of Schmidt and Hunter, such as Hunter, 1983a; Schmidt, Hunter, & Outerbridge, 1986; Schmidt, Hunter, Outerbridge, & Goff, 1988).

Given the empirical evidence about these three precursors of job performance, there appears to be little that would suggest higher job performance for males across a variety of jobs. For example, regarding declarative job knowledge, a study of post-training measures of job knowledge reported *d*s of approximately –.50 to –.60 across several clerical jobs in a military organization (Dunbar & Novick, 1988). Thus, females scored higher than males. Given that declarative or procedural knowledge is a precursor to job performance, this suggests that females might outperform males (although this is only one study).

There also appears to be little reason to posit gender differences in underlying or measured job performance due to motivation. In terms of trait motivation, differences between males and females on personality factors are relatively small. For example, Hough, Oswald, and Ployhart (2001) report a *d* of .06 for achievement motivation (see also Hyde, 2005). Similarly, there are few reported sizeable differences on other personality factors that might influence job performance to the advantage of males (e.g., Hough et al. report *d*s of .09 for extraversion, –.08 for overall conscientiousness, –.22 for dependability, .24 for adjustment, and –.39 for agreeableness; *N*s were more than 20,000 in all cases). Similarly, overall cognitive ability differences were generally minimal for males versus females (Hough et al., 2001).

Thus, females tend to score higher on the job knowledge variables and fairly similarly, on average, on other variables, such as trait motivation, personality, and cognitive ability. So, available data on precursors of job performance suggest that females might perform better than males on the job. Conversely, it also appears there is no reason to generally attribute higher underlying performance to males across a variety of jobs.

Promotability. In contrast to job performance ratings, promotability ratings are expected to favor males for a number of reasons. For example, females can receive less credit for their contributions (Hielman & Haynes, 2005). Likewise, female managers or leaders may have

to deal with more negative attitudes and may find that they have less access to leadership positions (e.g., see the review of role congruity theory by Eagly & Karau, 2002; see also Eagly et al., 1995).

EST suggests that promotion decisions may be more open to the influence of diffuse status cues because the process of considering the *predicted* level of performance in a higher level job involves more judgmental uncertainty than does the process of rating the *actual* performance of an incumbent in a given job. So, promotion decisions, by their nature, often do not allow decision makers maximal access to performance information (because most candidates have not done the job in question), thereby creating more opportunities for increased weight to diffuse cues.

Thus, we hypothesize:

Hypothesis 2: Males will score higher than females on promotability ratings.

Moderators. We also examined several moderators with exploratory analyses. For example, we coded time (i.e., date of study) to see if gender differences were relatively stable across decades due to factors such as the influence of specific cues (as per EST) or if *d*s changed across decades, as might be suggested by the potentially weakening of stereotypes or other related mechanisms (Nieva & Gutek, 1980). We also examined other potential moderators such as the purpose of ratings (as per Landy & Farr, 1980) to see if ratings for administrative purposes were associated with *d* values less than ratings made for research purposes (e.g., Tsui & O'Reilly, 1989). We also conducted similar analyses for job complexity and publication status (published vs. nonpublished documents).

Method

Literature Search

We searched databases for articles that reported data on measures of job performance effect size and gender. We searched PsycINFO (from the American Psychological Association), Business Source Premier, General Business File, Educational Resources Information Center (also known as ERIC), and Dissertation Abstracts International. We also checked the reference lists of previous analyses that focused on field data (e.g., Bowen et al., 2000). Finally, we wrote to researchers active in the area both for the availability of additional studies and to clear up any questions or get more detailed information.

Inclusion Criteria

There were four criteria necessary for studies to be included in our meta-analysis. First, studies must have reported measures of job performance (e.g., ratings, measures of output). As noted earlier, we did not include studies of salary (e.g., Ng et al., 2005) because we viewed salary as an outcome related to the assessment of job performance.

Second, ratings must have been completed by supervisors or peers. Thus, we did not include studies using self-report measures of job performance (e.g., Jagacinski, 1987), nor did we include student reports of teacher performance (as done by Bowen et al., 2000).

Third, participants being assessed must have been job incumbents. Studies asking students to rate paper or videotaped people were not included (as per studies in Davison & Burke, 2000; Olian et al., 1988).

Fourth, we paid careful attention to data dependence. We computed unit-weighted composites of performance when multiple correlations were present (e.g., Ployhart, Wiechman, Schmitt, Sacco, & Rogg, 2003). We also monitored data in order to avoid double counting similar or overlapping data sets published in different journals (see J. Wood's [2008] discussion of this issue). For example, we coded only one effect size for each of the following pairs of articles (e.g., Greenhaus & Parasuraman, 1993; Greenhaus, Parasuraman, & Wormley 1990; Pulakos & Wexley, 1983; Wexley & Pulakos, 1982). However, we did allow articles with multiple independent data sets to contribute multiple coefficients to our analyses (e.g., Arvey, Miller, Gould, & Burch, 1987; Saad & Sackett, 2002).

Coding

The first two authors, one male and one female, individually coded the data from each study. Measures of interrater agreement on the initial codes (correlation coefficients and percentage agreement) are reported below. Following the individual coding, the two coders met to resolve any discrepancies via consensus.

A number of variables were coded in addition to the sample size and the standardized gender group difference (respectively, interrater $r = .99$ and $r = 1.00$). [2] Coders first coded the nature of the measure (subjective vs. objective; 100%) and purpose of the measure (administrative, research, counseling, not sure; 96%). Second, the area of job performance was noted (e.g., rating of performance, output, quality or complaints, promotability; 100%). Job complexity was also coded using Hunter's (1983b) codes: low (e.g., mail sorter), low-medium (e.g., truck driver), medium (e.g., skilled crafts), medium-high (e.g., computer troubleshooter), or high (e.g., executives, scientists; $r = .92$). Finally, the date and publication status of the source was recorded (e.g., refereed journal, thesis or dissertation, etc.; 100%). We searched for the dates that data were gathered but found little such information, so we used date of publication for this variable when actual data collection dates were not available.

Analyses

We used the Hunter-Schmidt approach to meta-analysis (Hunder & Schmidt, 2004). Calculations were conducted with the Schmidt-Le program (version 1.1; Schmidt & Le, 2004). We corrected the effect sizes for unreliability in the measure of job performance. This generally involved the use of internal consistency measures of reliability from the performance data in our studies because there were only two studies that reported interrater reliabilities of ratings. Hunter and Schmidt state that using internal consistency measures of

Table 1
Meta-Analysis of Gender Differences for Measures of Overall Job Performance and Promotability

Analysis	Mean *d*	Corrected *d*	*K*	*N*	Var.	80% CRI	% Art.
Overall job performance	–.10	–.11	61	45,733	.030	–.33 to .12	17
Objective vs. subjective measures							
Supervisory ratings	–.13	–.14	50	36,896	.014	–.29 to .01	31
Objective measures	–.02	–.02	4	4,744	.015	–.18 to .14	20
Objective: no large sample	–.11	–.12	3	1,426	.040	–.37 to .14	20
Promotability ratings							
Promotability	.10	.11	8	4,550	.038	–.14 to .36	18
Matching studies promotability	.11	.12	6	4,307	.035	–.12 to .36	15
Matching studies performance	–.09	–.09	6	4,309	.001	–.12 to –.07	93
Manager studies							
Managers	–.14	–.15	8	7,555	.004	–.23 to –.07	55
Managers: no large samples	–.08	–.09	6	1,961	.005	–.18 to .002	74

Note: Mean *d* is the observed *d* across studies, corrected *d* is corrected for measurement reliability in job performance, *K* is the number of coefficients, *N* is the number of participants, Var. is the variance of the estimate of the corrected *d,* CRI is an 80% credibility interval, and % Art. is the percentage of variance due to the artifacts of sampling error and measurement unreliability. We also removed police and military studies from these analyses, given their somewhat different pattern of results from other studies (details available from the first author).

reliability is generally expected to produce a conservative correction of effect sizes. For some of our ancillary analyses, we used the reliability value of .6 for measures of promotability. We adopted this value from previous literature (e.g., McKay & McDaniel, 2006; Roth et al., 2003). This value is also similar to the reliability of supervisor ratings of leadership ability ($r = .53$, $K = 10$, $N = 2{,}171$) in the meta-analysis by Viswesvaran, Ones, and Schmidt (1996).

Results

Overall Results

The results for male versus female differences for measures of job performance are reported in Table 1. Our overall analyses include studies using performance ratings and measures of output (e.g., sales volume), although only three studies in this analysis contained objective measures of performance.[3] We do *not* include our measures of promotion potential in these overall analyses, as they are noted separately in the table.

The overall analysis of job performance measures resulted in a mean corrected *d* of –.11 ($K = 61$, $N = 45{,}733$). This suggests that females, on average, were rated as performing somewhat better than males in operational field settings. The 80% credibility interval ranged from –.33 to .12. This suggests that there is substantial variation in estimated population values and that in some cases males perform better (positive values) and that in a somewhat

greater proportion of cases females perform better (negative values). In general, the data show substantial gender similarity in mean levels of overall job performance.

Overall, the data support our hypothesis (Hypothesis 1) of relatively small differences in measures of operational job performance (and if there are differences, as indicated by the range in credibility values, females generally have slightly higher average ratings or output). The exclusion of large samples does not substantially change the mean estimates.[4]

We also report results of supervisory or subjective ratings versus objective measures (again, see Table 1). Results for supervisory ratings indicate that $d = -.14$ ($K = 50$, $N = 36{,}896$), with a credibility interval of –.29 to .01. Objective measures are associated with $d = -.02$ ($K = 4$, $N = 4{,}744$). Results change somewhat when a large sample is deleted, to $d = -.12$ ($K = 3$, $N = 1{,}426$). We interpret the latter values with caution because the number of studies and sample sizes are small. Results for objective measures paint a reasonably consistent picture with performance ratings in which gender differences were small (although females score slightly higher on average). Thus, there is some convergence across different types of measures.

Promotability. Several studies reported measures of promotability (e.g., Harris, Kacmar, & Carlson, 2006). As noted in Table 1, the corrected mean d was .11 ($K = 8$, $N = 4{,}550$). Thus, on average, males scored somewhat higher than females on measures of promotability, although the credibility interval was again wide. Interestingly, a subset of six studies also reported measures of performance on the current job (*N*s differ slightly on these analyses due to missing data in one primary study). Results show a mean d of –.09 for job performance ($K = 6$, $N = 4{,}309$) such that female performance was slightly higher, while at the same time the mean d for ratings of promotability was .12 ($K = 6$, $N = 4{,}307$). Below, we return to this unique finding and analysis (directly comparing the promotability d to the performance *d)*.

We also report results for job performance for manager samples separately, given past attention to this type of job (Eagly et al., 1992; see also Eagly & Karau, 2002, or Schein, 2001). In the bottom panel of Table 1, we also report a mean corrected d of –.15 ($K = 8$, $N = 7{,}555$) for all managerial samples, and we note that the credibility interval does not include zero. The mean d is –.09 ($K = 6$, $N = 1{,}961$) after removing one large sample, and the credibility interval does include zero. Thus, point estimates of gender group differences (i.e., mean *d*s) suggest more similarity than difference, although sample sizes are small. Additionally, consistent with our overall analyses, the tendency is that for nonzero values the female average is greater than the male average. We also note that these jobs were more representative of middle management positions than of upper management positions.

Moderators

Before examining some of our other moderators, we examined the type of organization associated with the data (see Table 2), based on strong statements in previous meta-analyses. Specifically, Eagly et al. (1995) suggested that "military organizations deviated strongly from all other classes of studies" (p. 135; see p. 138 for further cautions).

Table 2
Meta-Analysis of Time, Publication Status, Job Complexity, and Purpose of Data Gathering

Analysis	Mean *d*	Corrected *d*	*K*	*N*	Var.	80% CRI	% Art.
Type of organization							
Industrial	–.12	–.12	41	36,326	.013	–.27 to .02	29
Public sector and educational	–.31	–.33	8	1,147	.045	–.61 to –.07	42
Public sector: no large sample	–.19	–.20	7	769	.019	–.38 to –.02	68
Military/police	.00	.00	12	8,260	.086	–.37 to .37	7
Military/police: no large sample	.06	.07	11	5,276	.123	–.38 to .51	7
Date of publication							
1969-1988	–.14	–.15	18	24,231	.007	–.25 to –.04	34
1969-1988: no large samples	–.18	–.19	16	7,675	.011	–.33 to –.06	46
1989-2001	–.11	–.12	14	3,315	.031	–.34 to .10	39
1989-2001: no large samples	–.14	–.15	13	2,521	.039	–.45 to –.01	49
2002-2009	–.05	–.05	29	18,187	.055	–.36 to .25	11
2002-2009: no large samples	–.03	–.03	27	12,259	.075	–.38 to .32	12
Publication status							
Journal	–.10	–.11	44	26,843	.041	–.37 to .15	15
Not published	–.11	–.12	17	18,872	.012	–.26 to .02	26
Not published: no large samples	–.08	–.09	15	2,316	.071	–.43 to .25	30
Job complexity							
Low and low-medium complexity	–.07	–.07	29	19,404	.049	–.36 to .22	12
Medium complexity	–.12	–.13	14	15,645	.014	–.28 to .02	23
Medium-high complexity	–.13	–.13	9	8,841	.003	–.20 to –.07	64
Medium-high complexity: no large samples	–.06	–.07	7	3,247	.000	–.07 to –.07	100
Purpose of performance measure							
Administrative	–.06	–.07	12	7,341	.024	–.27 to .13	23
Administrative: no large samples	–.13	–.14	11	4,023	.034	–.38 to .10	27
Research	–.10	–.11	43	35,411	.031	–.33 to .12	15
Research: no large samples	–.07	–.08	39	13,434	.077	–.43 to .28	15

Note: Mean *d* is the observed *d* across studies, corrected *d* is corrected for measurement reliability in job performance, *K* is the number of coefficients, *N* is the number of participants, Var. is the variance of the estimate of the corrected *d,* CRI is an 80% credibility interval, and % Art. is the percentage of variance due to the artifacts of sampling error and measurement unreliability.

In our database, industrial samples were associated with a mean corrected *d* of –.12 (*K* = 41, *N* = 36,326), variance was .013, and the credibility interval ranged from –.27 to .02. Similarly, public-sector and educational jobs were associated with a mean corrected *d* of –.20 (*K* = 7, *N* = 769) and a variance of .019 when a large sample was removed. In contrast, military and police jobs were associated with a mean corrected *d* of .00 (*K* = 12, *N* = 8,260) and a mean corrected *d* of .07 (*K* = 11, *N* = 5,271) when one large sample was removed. Variability was relatively high (variance was .086, or .123 when the large sample was removed), approximately 3.5 to 10 times as much variance as in other studies. Overall, military studies may have a different pattern of means and variances than do other job categories.

Table 2 also presents our results for date of publication. We grouped studies into three time periods that were approximately equal in terms of the number of studies. Studies from

1969 to 1989 were associated with a mean corrected $d = -.15$. The mean d for the period from 1989 to 2001 was –.12, while the mean d for studies from 2002 to 2009 was –.05. Thus, there is some evidence that male and female differences in performance evaluations may be decreasing, although credibility intervals do overlap. A similar pattern was obtained across the three time periods when large samples were removed. A reviewer suggested that this pattern of results might be consistent with an increasing number of capable females entering the workforce as time passed.

Results in Table 2 suggest little in the way of moderation for publication status. The corrected mean d for data from journals is –.11, while the mean d associated with unpublished sources (i.e., dissertations, theses, and convention presentations) is –.12.

There is no clear trend regarding the moderator of job complexity (again, see Table 2). Lower complexity jobs were associated with a mean corrected d of –.07. Medium complexity jobs and medium-high complexity jobs were both associated with a corrected d of –.13 (and removing two large medium-high complexity samples resulted in a d of –.07).

Results in Table 2 also do not suggest a clear pattern of moderation for administrative versus research measures of performance. Administrative performance differences are associated with a mean d of –.07, while research measures are associated with a mean d of –.11. However, the pattern reverses itself when large samples are removed. For publication status, job complexity, and research purpose moderators, the credibility intervals overlap considerably.

Discussion

There are multiple views in the literature regarding the magnitude (and direction) of gender differences in organizational measures of job performance. Some research suggests that males may, on average, receive higher performance ratings than females receive (e.g., Nieva & Gutek, 1980, as well as Greenhaus & Parasuraman, 1993). Yet, EST suggests that specific status cues regarding job-relevant information (which can also be conceptualized as individuating information) can be more important than diffuse status cues in judgments of job performance. Further, the gender similarities hypothesis (Hyde, 2005) stresses the lack of differences on many psychological individual difference variables that act as theoretical antecedents of job performance.

Overall analyses. We conducted a meta-analysis to investigate gender-based differences on measures of job performance in field studies. In contrast to some previous suggestions (e.g., Nieva & Gutek, 1980), we found that mean gender differences in measures of job performance were generally what one might consider small (i.e., standardized differences smaller than approximately .20 in magnitude; cf. Murphy & Myors, 2004, for a discussion of effect sizes). That is, mean performance levels for males and females are generally more similar than different. If anything, on average, females tended to slightly outscore males. Thus, consistent with EST, data from our meta-analysis support gender similarity for the variable of job performance.

To our knowledge, the current analyses are the most extensive and clearest meta-analytic estimates of gender differences available when on-the-job performance is the focal construct. That is, other meta-analyses examined constructs such as hiring suitability in laboratory studies (e.g., the overall mean *d* of .41 from Olian et al., 1988). Another meta-analysis focused on the construct of leader performance rather than job performance (Eagly et al., 1995). This latter study also aggregated across laboratory studies, field studies, and a heterogeneous mix of dependent measures (e.g., performance ratings by others, performance ratings by self, satisfaction with the leader, etc.). Thus, we conducted our analysis to provide a clearer picture of gender differences on the construct of operational job performance.

Our results are also different from those of Sackett et al.'s (1991) large primary study in which $d = .07$ (for many blue-collar jobs).[5] Our more generalizable, meta-analytic difference of –.11 is in a different direction (i.e., has a different sign), suggesting that if there is a difference in performance, females generally receive higher performance ratings and/or perform higher. Further, our study cumulates results across approximately 60 samples from a variety of types of jobs and samples.

Also, given that many of our meta-analytic *d*s were negative, it appears that using performance ratings as a complete or partial predictor of future organizational behavior is unlikely to lead to adverse impact for females (i.e., hiring a smaller proportion of women) in many jobs. However, this may not necessarily be the case for some jobs with a male stereotype (Davison & Burke, 2000).

Overall, our data suggest that males' and females' mean measured levels of job performance are highly similar, and *if* there is a difference, females have a slight advantage. Thus, Hypothesis 1 was supported.

Promotion potential analyses. Our analyses of promotion potential suggest that females are rated lower than males. That is, for the six matched studies, females had slightly better current job performance ratings than did males ($d = –.09$) but *at the same time* had slightly lower promotability ratings ($d = .12$). Thus, Hypothesis 2 was supported.

Taken together, our contrasting results for Hypothesis 1 and Hypothesis 2 are interesting. First, we note that these results are consistent with EST; that is, specific cues (individuating information) are more readily available when rating current performance, but when rating promotion potential, the lack of specific information gives diffuse cues relatively more room to operate. These results could also be interpreted as supportive of the notion in role congruity theory that posits that females may have less access to higher level jobs via the promotion process (Eagly & Karau, 2002). Our empirical results could also be seen as consistent with the lack-of-fit model (Heilman, 1983; Lyness & Heilman, 2006) and concerns that females are perceived by managers as having more work–life conflict (Hoobler, Wayne, & Lemmon, 2009). Again, the increased ambiguity for predicting performance on a new job may be associated with increased use of stereotypes or diffuse cues versus ratings of performance on an existing job (see also Eagly et al., 1995). We urge substantial future research targeted at this finding. We suggest that the dynamic of greater ambiguity of predicting future job performance may also occur during the personnel hiring or selection process (more on this below).

Regarding the concept of gender bias, the number of negative (and near-zero) meta-analytic *d*s suggests that relatively straightforward theoretical views (i.e., bias overpowers qualifications) do not necessarily apply to the assessment of current, operational job performance. Similarly, although our managerial sample was small, female managers were not rated lower on performance than male managers. Thus, our job performance results are not generally supportive of role incongruity theories for the variable of job performance (as per Eagly & Karau, 2002), although we note that our managers were more middle-level managers than upper-level managers. Again, a theoretical explanation for our findings is that specific status cues (i.e., job-related information) are present when supervisors rate subordinates. Other individuals might call this individuating information (e.g., Eagly et al., 1995; Eagly et al., 1992; Reskin, 2000).

Our results do not directly address more complex and subtle theories, such as double standards in which (a) there are higher standards invoked for lower status individuals to get a high evaluation (i.e., relatively easier standards are used for higher status individuals to get a high evaluation) and (b) at the same time, lower standards are invoked for lower status individuals when giving evaluations of minimal competence (i.e., relatively harder standards are used for higher status individuals to be deemed minimally competent; Biernat & Manis, 1994; see also Maurer & Taylor, 1994). Such theories deserve more theoretical and empirical attention. Similarly, the delineation of staff versus line jobs (and associated beliefs and stereotypes) may also be an important distinction in understanding performance assessment and bias (e.g., Lyness & Heilman, 2006).

A reviewer also suggested the possibility of bias against females despite the reported similar job performance ratings. That is, females have some noted advantages in precursors of job performance, yet the *d*s for job performance ratings are close to zero. In particular, we noted that females tend to score higher in a series of samples on measures of declarative job knowledge in one military study and that females, on average, have many similar levels of job-related personality traits. Further, females tend to have higher levels of social skills, as assessed by self-report measures of samples of college students. Studies by Riggio and colleagues (e.g., Riggio, 1986; Riggio, Tucker, & Coffaro, 1989) report *d*s of –.23 and –.46 in two samples. Some meta-analytic work suggests the advantage by females is even higher in terms of "reading" nonverbal social information, as *d*s sometimes approach –1.0 when a wide variety of verbal and nonverbal information is available to individuals interpreting social skills (Hall, 1978; see also Hall, Carter, & Horgan, 2000). In sum, the reviewer noted that, although *d*s are relatively small in magnitude, bias might exist against females because the bias decreases *d*s to their current levels that ought to be higher given trends in antecedents of job performance.

We suggest future research on this complex interplay of variables. Such research could lead to important insights, although it would likely require a well-specified model of job performance for a given job. Such models would require substantial care, as a complex array of gender-related abilities might be related to job performance (e.g., spatial ability, muscular endurance, agreeableness or dependability; cf. Hough et al., 2001). And researchers should note that job performance *d*s will not necessarily be as large as precursor *d*s because of regression effects (however, to the extent that precursor *d*s are positive, one could expect that job performance *d*s are also positive). Further, one would need to attend to the intercorrelations of

abilities to carefully investigate how the antecedents of performance and bias would work in concert to influence measures of job performance.

Limitations

All research efforts have limitations, and we note several. First, although our meta-analytic sample size was $K = 61$ studies (for performance ratings and measures of output), use of moderator analyses led to relatively small numbers of coefficients (K) for some analyses (e.g., promotability).

Second, similar to other researchers (Sackett et al., 1991), we could not readily analyze subdimensions of performance. While we looked for this information, it was generally not possible to accurately analyze particular parts of the task performance domain (e.g., interacting with others vs. technical performance).

Third, we did not have access to "true scores" for job performance. Nor were we able to find studies such that all factors other than gender were controlled. Thus, given our desire to focus on field studies, we did not directly address ratings bias (as might be possible in the laboratory).

Future Research

We suggest that future research address different dimensions of job performance. One approach is to use the gender roles perspective to make predictions about performance differences. This perspective suggests that females are more likely to learn to deal with the world in a social or communal way. Males may have a more agentic approach that emphasizes more task-oriented behaviors (Eagly & Johannesen-Schmidt, 2001). These emphases may give rise to different gender-related skills and different levels of performance on factors such as interpersonal tasks versus technical tasks—which might vary in emphasis across different jobs.

In particular, research might pursue gender differences within work sample measures of job performance. Work samples, as measures of job performance, have the advantage of being more standardized in terms of opportunities for observation and scoring than are many supervisory ratings of job performance. They can also be designed to focus on particular skill sets, so they could foster a more construct-oriented approach to examining gender performance differences. Unfortunately, we found no work samples that were conceptualized as measures of job performance in our gender-based literature search. Similar work might also be applied to assessment centers, and "head-to-head" tests of gender group differences on work samples, assessment centers, ratings of job performance, and ratings of promotability could be conducted.

We also suggest more study of measures of promotion potential. Again, we found only eight such studies. Researchers might also pursue meta-analyses of withdrawal behaviors (e.g., lateness, absenteeism, turnover).

We also encourage work on extra-role behavior or organizational citizenship behaviors (OCBs; Borman, White, Pulakos, & Oppler, 1991; Vey & Campbell, 2004). Future research

could focus on issues such as the context or situation and on the expectations of those evaluating performance. For example, various dimensions of OCBs might be expected more for men or women (e.g., altruism and courtesy dimensions are considered "in role" behavior more for women than for men; cf. Kidder & McLean-Parks, 2001). Likewise, it is possible that withholding altruistic behavior may be more harmful to evaluations of women than of men (Heilman & Chan, 2005).

We also suggest continuing research into gender differences in hiring. Theoretically, hiring might be viewed as similar to the promotion process. That is, the prediction of future performance might allow more diffuse status cues to influence pre-employment-related judgments because there can be less individuating information related to the quality of the applicants. While there has been substantial work on gender differences on objective tests of variables such as cognitive ability (e.g., Hough et al., 2001), there is less research on more subjective predictors of job performance such as application blanks, interviews, ratings in work sample tests, and so on. Again, we suggest such work use field samples of job applicants whenever possible.

It would also be useful to conduct more research on multiple, potentially interacting, demographic variables. For example, as implied by a reviewer, it would be interesting to investigate if gender and ethnicity combine to influence performance assessments. Similarly, the joint role of age and gender, or age and ethnicity, on job performance might reveal how diffuse status cues interrelate to stereotypes. Such research designs could be used to study other dependent variables such as hiring, promotion, and salary.

Finally, we suggest research into gender differences in political skills. One researcher found that although females performed better on the job than males did, the females were rated as having less influence and less centrality to the dominant coalition of key decision makers (Brass, 1985). This could limit or influence perceptions of promotion potential. Other researchers have suggested that females may have lower levels of organizational political skills, thereby reducing the opportunity to interact with key decision makers and decreasing their chances of promotion (Ferris, Frink, Galang, Zhou, & Howard, 1996; Perrewe & Nelson, 2004).

In sum, we conducted a meta-analysis of gender differences in job performance in field studies. Our findings are generally consistent with predictions of small differences. Females scored slightly higher than males, although credibility intervals often included zero. On the other hand, ratings of promotion potential tended to favor males. Thus, we suggest that ratings of promotability deserve further research attention, that results are consistent with EST, and that future research should consider more complex models regarding any gender bias and job performance differences. We look forward to the knowledge and continued insights that such research will bring.

Notes

1. We used the *d* statistic (i.e., a standardized, mean gender group difference) to provide an index of group differences. The numerator of the *d* statistic is the mean male rating minus the mean female rating. The denominator is the pooled or averaged value of the male and female group standard deviations. Thus, a *d* of –.25 would indicate that the mean male score is one quarter of an averaged standard deviation lower than the mean female score.

2. There was one influential case in coding effect sizes in which one coder mistakenly reversed the sign on the effect size. We trimmed this case (Lipsey & Wilson, 2001) from our analyses of agreement. Interrater agreement was .59 with the case and 1.00 without the case.

3. One study reported both an objective and a subjective measure of job performance. We used the subjective measure in overall analyses, given its similarity to most other measures of job performance in this analysis.

4. There were several large samples ($N > 2{,}000$) that might have influenced results in this analysis (and throughout the rest of the article). In the text, we generally discuss results for all samples unless the large samples have a marked influence on results.

5. We were unable to incorporate Sackett, Dubois, and Noe's (1991) data into our analysis, given the summary form (e.g., frequency distribution) in which it was reported.

References

Asterisks denote studies used in our meta-analysis.

*Abraham, J. D., & Morrison, J. D. 2003. Relationship between the performance perspectives inventory's implementing scale and job performance of incumbents in a dual security/service role. *Applied HRM Research,* 8(2): 73-76.

*Arvey, R. D., Miller, H. E., Gould, R., & Burch, P. 1987. Interview validity for selecting sales clerks. *Personnel Psychology,* 40: 1-12.

*Barrick, M. R., Mount, M. D., & Strauss, J. P. 1994. Antecedents of involuntary turnover due to a reduction in force. *Personnel Psychology,* 47: 515-535.

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

Berger, J., Rosenholtz, S. J., & Zelditch, M. 1980. Status organizing processes. *Annual Review of Sociology,* 6: 479-508.

Berger, J., Wagner, D. G., & Zelditch, M. 1985. Introduction: Expectation states theory. In J. Berger & M. Zelditch (Eds.), *Status, rewards, and influence:* 1-72. San Francisco: Jossey-Bass.

Biernat, M., & Manis, M. 1994. Shifting standards and stereotype based judgments. *Journal of Personality and Social Psychology,* 66: 5-20.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. 1991. Models of supervisory job performance ratings. *Journal of Applied Psychology,* 76: 863-872.

Bowen, C. C., Swim, J. K., & Jacobs, R. R. 2000. Evaluating gender biases on actual job performance of real people: A meta-analysis. *Journal of Applied Social Psychology,* 30: 2194-2215.

*Brass, D. J. 1985. Men's and women's networks: A study of interaction patterns and influence in an organization. *Academy of Management Journal,* 28: 327-343.

*Byron, K. 2007. Male and female managers' ability to "read" emotions: Relationships with supervisor's performance ratings and subordinates' satisfaction ratings. *Journal of Organizational and Occupational Psychology,* 80: 713-717.

Campbell, J. P., Gasser, M. B., & Oswald, F. L. 1996. The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations:* 258-299. Hillsdale, NJ: Lawrence Erlbaum.

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:* 35-70. San Francisco: Jossey-Bass.

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

Cleary, T. A. 1968. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement,* 5: 115-124.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. 2001. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology,* 86: 425-445.

*Confidential Industry Technical Report.

*Confidential Industry Technical Report.

*Confidential Industry Technical Report.

*Cook, C. D. 1984. *The validation of an assessment center procedure for the early identification of supervisory job performance and potential for advancement.* Unpublished doctoral dissertation, University of Wisconsin–Madison.

Correll, S. J., & Ridgeway, C. L. 2003. Expectation states theory. In J. Delamater (Ed.), *Handbook of social psychology:* 29-51. New York: Kluwer Academic/Plenum.

*Crane, J. B. 1981. *The interrelationship of self-concept and performance and selected characteristics of extension 4-H youth agents within the Mississippi cooperative extension service.* Unpublished doctoral dissertation, Mississippi State.

*Cropanzano, R., Rupp, D. E., & Zinta, S. B. 2003. The relationship of emotional exhaustion to work attitudes, job performance, and organizational citizenship behaviors. *Journal of Applied Psychology,* 88: 160-169.

Davison, H. K., & Burke, M. J. 2000. Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior,* 56: 225-248.

*Dereinda, R. 1984. *The relationships between job satisfaction and job performance among county agents in the Wisconsin cooperative extension service*. Unpublished doctoral dissertation, University of Wisconsin–Madison.

*Diefendorff, J. M., Brown, D. J., Kamin, A. M., & Lord, R. G. 2002. Examining the roles of job involvement and work centrality in predicting organizational citizenship behaviors and job performance. *Journal of Organizational Behavior,* 23: 93-108.

Dovidio, J. F., Ellyson, S. L., & Keating, C. F. 1988. The relationship of social power to visual displays of dominance between men and women. *Journal of Personality and Social Psychology,* 54: 233-242.

Dunbar, S. R., & Novick, M. R. 1988. On predicting success in training for men and women: Examples from the Marine Corps clerical specialties. *Journal of Applied Psychology,* 73: 545-550.

Eagly, A. H., & Johannesen-Schmidt, M. C. 2001. The leadership styles of women and men. *Journal of Social Issues,* 57: 781-797.

Eagly, A. H., & Karau, S. J. 2002. Role congruity theory of prejudice toward female leaders. *Psychological Review,* 109: 573-598.

Eagly, A. H., Karau, S. J., & Makhijani, M. G. 1995. Gender and effectiveness of leaders: A meta-analysis. *Psychological Bulletin,* 117: 125-145.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. 1992. Gender and evaluation of leaders: A meta-analysis. *Psychological Bulletin,* 111: 3-22.

Farh, J., & Dobbins, G. H. 1989. Effects of self-esteem on leniency bias in self-reports of performance: A structural equation model analysis. *Personnel Psychology,* 42: 835-850.

*Ferris, G. R., Bergin, T. G., & Wayne, S. J. 1988. Personal characteristics, job performance, and absenteeism of public school teachers. *Journal of Applied Social Psychology,* 18: 552-563.

Ferris, G. R., Frink, D. D., & Galang, M. C., Zhou, J., & Howard, J. L. 1996. Perceptions of organizational politics: Predictions, stress related implications, and outcomes. *Human Relations,* 49: 233-266.

Freese, L., & Cohen, B. P. 1973. Eliminating status generalization. *Sociometry,* 36: 177-193.

*Glascock, L. B. 1995. *Attributional style as a predictor of work performance.* Unpublished doctoral dissertation, George Washington University.

*Green, C., Jegadeesh, N., & Tang, Y. 2009. Gender and job performance: Evidence from Wall Street. *Financial Analysts Journal,* 65: 1-14.

Greenhaus, J. H., & Parasuraman, S. 1993. Job performance attributions and career advancement prospects: An examination of gender and race effects. *Organizational Behavior and Human Decision Processes,* 55: 273-287.

*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, J. A. 1978. Gender differences in decoding nonverbal cues. *Psychological Bulletin,* 85: 845-857.

Hall, J. A., Carter, J. D., & Horgan, T. G. 2000. Gender differences in nonverbal communication of emotion. In A. Fischer (Ed.), *Gender and emotion: Social psychological perspectives:* 97-117. Cambridge, UK: Cambridge University Press.

*Harris, K. J., Kacmar, K. M., & Carlson, D. S. 2006. An examination of temporal variables and relationship quality on promotability ratings. *Group and Organization Management,* 31: 677-699.

Harrison, D., Newman, D., & Roth, P. L. 2006. How important are job attitudes? Meta-analytic comparisons of integrative behavioral outcomes and time sequences. *Academy of Management Journal,* 49: 305-326.

Harrison, D. A., & Shaffer, M. A. 1994. Comparative examinations of self-reports and perceived absenteeism norms: Wading through Lake Wobegon. *Journal of Applied Psychology,* 79: 240-251.

Heilman, M. E. 1983. Sex bias in work settings: The lack of fit model. *Research in Organizational Behavior,* 5: 269-298.
Heilman, M. E. 2001. Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues,* 57: 657-674.
Heilman, M. E., & Chen, J. J. 2005. Same behavior, different consequences: Reactions to men's and women's altruistic citizenship behavior. *Journal of Applied Psychology,* 90: 431-441.
Heilman, M. E., & Haynes, M. C. 2005. No credit where credit is due: Attributional rationalization of women's success in male-female teams. *Journal of Applied Psychology,* 90: 905-916.
*Helmick, T. A. 1987. *The effect of gender, job classification and self-esteem upon job performance.* Unpublished master's thesis, Central Missouri State University.
*Hoobler, J. M., Wayne, S., & Lemmon, G. 2009. Bosses' perceptions of family-work conflict and women's promotability: Glass ceiling effects. *Academy of Management Journal,* 52: 939-957.
Hough, L. M., Oswald, F. L., & Ployhart, R. E. 2001. Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment,* 9: 152-194.
Hunter, J. E. 1983a. A causal analysis of cognitive ability, job knowledge, and job performance, and supervisory ratings. In R. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory:* 257-266. Hillsdale, NJ: Lawrence Erlbaum.
Hunter, J. E. 1983b. *Test validation for 13,000 jobs: An application of job classification and validity generalization for the General Aptitude Test Battery (GATB)*. USES Test Research Rep. No. 450. Washington, DC: U.S. Department of Labor.
Hunter, J. E., & Schmidt, F. L. 2004. *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
Hyde, J. S. 2005. The gender similarities hypothesis. *American Psychologist,* 60: 581-592.
*Igbaria, M., & Baroudi, J. J. 1995. The impact of job performance evaluations on career advancement prospects: An examination of gender differences in the IS workplace. *MIS Quarterly,* 19(1): 107-123.
*Jepsen, D. M., & Rodwell, J. J. 2009. Justice in the workplace: The centrality of social versus judgmental predictors of performance varies by gender. *International Journal of Human Resource Management,* 20: 2066-2083.
Jagacinski, C. M. 1987. Androgyny in a male-dominated field: The relationship of sex-typed traits to performance and satisfaction in engineering. *Sex Roles,* 17: 529-547.
*Joerger, J. H. 1984. *A field study of fear of success, internal-external locus of control and performance ratings among professional men and women.* Unpublished doctoral dissertation, University of Northern Colorado.
*Joshi, A., Liao, J., & Jackson, S. 2006. Cross-level effects of workplace diversity on sales performance and pay. *Academy of Management Journal,* 49: 459-481.
Kidder, D. L., & McLean Parks, J. 2001. The good soldier: Who is s(he)? *Journal of Organizational Behavior,* 22: 939-959.
King, E. B., Hebl., M. R., George, J. M., & Matusik, S. F. (2010). Understanding tokenism: Antecedents and consequences of a psychological climate of gender inequity. *Journal of Management,* 36: 482-510.
Landy, F. J., & Farr, J. L. 1980. Performance rating. *Psychological Bulletin,* 87: 72-107.
Lipsey, M. W., & Wilson, D. B. 2001. *Practical meta-analysis*. Thousand Oaks, CA: Sage.
*Lyness, K. S., & Heilman, M. E. 2006. When fit is fundamental: Performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology,* 91: 777-783.
MacGillivray, E. D., Beecher, H. J. M., & Golden, D. 2009. Legal developments—Gender issues now, and increasingly in the US and EU spotlight. *Global Business and Organizational Excellence,* 28(3): 79-86.
*Marion-Landais, C. A. 1999. *Dyadic agreement about ideal leadership behaviors and its relationship to job satisfaction and performance ratings.* Unpublished doctoral dissertation, University of South Florida.
Maurer, T. M., & Taylor, M. A. (1994). Is sex by itself enough: An exploration of gender bias issues in performance appraisal. *Organizational Behavior and Human Decision Processes,* 60: 231-251.
McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology,* 79: 493-505.
McKay, P. F. 2009. Perspectives on adverse impact in work performance: What we know and what we could learn more about. In J. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection:* 249-270. New York: Routledge.

McKay, P. F., & McDaniel, M. A. 2006. A reexamination of Black–White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology,* 91: 538-554.

*Michael, D. F. 2004. *Gender differences in information-exchange and outcomes at work*. Unpublished doctoral dissertation, Auburn University.

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

Mintzberg, H. 2001. The yin and yang of managing. *Organizational Dynamics,* 29: 306-312.

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

*Morrison, J. D., & Abraham, J. D. 2003. Relationship between the performance perspectives inventory's selling scale and job performance of real estate agents. *Applied HRM Research,* 8(2): 85-88.

*Motowildo, S. J., & Schmit, M. J. 1997. *Performance assessment and interview procedures for store manager and store associate positions* (Final tech. rep.). University of Florida, Human Resource Research Center.

Murphy, K. R., & Myors, B. 2004. *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Ng, T. W. H., Eby, L., Sorensen, K. L., & Feldman, D. C. 2005. Predictors of objective and subjective career success. A meta-analysis. *Personnel Psychology,* 58: 367-408.

Nieva, V. F., & Gutek, B A. 1980. Sex effects on evaluation**.** *Academy of Management Review,* 5: 267-276.

*O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L., & Lawrence, A. 2007. Incremental validity of situational judgment tests for task and contextual performance. *International Journal of Selection and Assessment,* 15: 19-29.

Olian, J. D., Schwab, D. P., & Haberfield, Y. 1988. The impact of applicant gender compared to qualifications on hiring recommendations: A meta-analysis of experimental studies. *Organizational Behavior and Human Decision Processes,* 41: 180-195.

*Ostroff, C., Atwater, L. E., & Feinberg, B. J. 2004. Understanding self–other agreement: A look at rater and rater characteristics, context, and outcomes. *Personnel Psychology,* 57: 333-375.

*Patterson, T. F., Jr. 1984. A *study to determine the relationship between rural New England extension agent educational orientation and job performance*. Unpublished doctoral dissertation, Indiana University.

*Pazy, A., & Oron, I. 2001. Sex proportion and performance evaluation among high-ranking military officers. *Journal of Organizational Behavior,* 22: 689-702.

Perrewe, P. L., & Nelson, D. L. 2004. The facilitative role of political skill. *Organizational Dynamics,* 33: 366-378.

*Phillips, J. F. 1992. Predicting sales skills. *Journal of Business and Psychology,* 7: 151-160.

*Phillips, J. F. 1993. Predicting negotiation skills. *Journal of Business and Psychology,* 7: 403-411.

*Plamondon, K. E., & Schmitt, N. 2000. *Validity and subgroup differences of combinations of predictors as a function of research design.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, New Orleans.

*Ployhart, R. E., Wiechmann, D., Schmitt, N., Sacco, J. M., & Rogg, K. 2003. The cross-cultural equivalence of job performance ratings. *Human Performance,* 16: 49-79.

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

*Pulakos, E. D., & Wexley, K. N. 1983. The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. *Academy of Management Journal,* 26: 129-139.

*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.

Reskin, B. F. 2000. The proximate causes of employment discrimination. *Contemporary Sociology,* 29: 319-328.

Riggio, R. E. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology,* 51: 649-660.

Riggio, R. E., Tucker, J., & Coffaro, D. 1989. Social skills and empathy. *Personality and Individual Differences,* 11: 93-99.

Robertson, T. M. 1986. Fundamental strategies for wage and salary administration. *Personnel Journal,* 65(11): 120-132.

Roth, P. L., Huffcutt, A. I., & Bobko, P. 2003. Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology,* 88: 725-740.

*Saad, S., & Sackett, P. R. 2002. Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology,* 87: 667-674.

Sackett, P. R., Dubois, C. L. Z., & Noe, A. W. 1991. Tokenism in performance evaluation: The effects of work group representation on male–female and White–Black differences in performance ratings. *Journal of Applied Psychology,* 76: 263-267.

Schein, V. E. 2001. A global look at psychological barriers to women's progress in management. *Journal of Social Issues,* 57: 675-688.

*Schmidli, K. R. 1969. *Effectiveness of teachers of the educable mentally retarded as related to preparation*. Unpublished doctoral dissertation, Colorado State College.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. 1986. The 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., Hunter, J. E., Outerbridge, A. N., & Goff, S. 1988. Joint relation of experience and ability with job performance: Test of three hypotheses. *Journal of Applied Psychology,* 73: 46-57.

Schmidt, F. L., & Le, H. 2004. *Software for the Hunter–Schmidt meta-analysis methods*. Department of Management and Organization, University of Iowa, Iowa City.

*Schmit, M. J., & Motowildo, S. J. 1995. *Development and validation of a situational interview and personality test for selecting sales associates* (Final tech. rep.). University of Florida, Human Resource Research Center.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. 2007. An integrative model of organizational trust: Past, present, and future. *Academy of Management Review,* 32: 344-354.

*Shore, L. M., & Thornton, G. C. 1986. Effects of gender on self- and supervisory ratings. *Academy of Management Journal,* 29: 115-129.

*Shore, T. H. 1992. Subtle gender bias in the assessment of managerial potential. *Sex Roles,* 27: 499-515.

Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. 1989. Joan McKay versus John McKay: Go gender stereotypes bias evaluations. *Psychological Bulletin,* 105: 409-429.

*Thacker, R. A., & Wayne, S. J. 1995. An examination of the relationship between upward influence tactics and assessments of promotability. *Journal of Management,* 21: 739-756.

*Tsui, A. S., & Gutek, B. A. 1984. A role set analysis of gender differences in performance, affective relationships, and career success of industrial middle managers. *Academy of Management Journal,* 27: 619-635.

Tsui, A. S., & O'Reilly, C. A. 1989. Beyond simple demographic effects: The importance of relational demography in superior-subordinate dyads. *Academy of Management Journal,* 32: 402-423.

Vey, M. A., & Campbell, J. P. 2004. In-role or extra-role organizational citizenship behavior: Which are we measuring? *Human Performance,* 17: 119-135.

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.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. 2005. Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology,* 90: 108-131.

*Wanous, J. P., Stumpf, S. A., & Bedroisan, H. 1979. Job survival of new employees. *Personnel Psychology,* 32: 651-662.

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

*Wexley, K. N., & Pulakos, E. D. 1982. Sex effects on performance ratings in manager–subordinate dyads: A field study. *Journal of Applied Psychology,* 67: 433-439.

Wood, J. 2008. Methodology for dealing with duplicate effects in a meta-analysis. *Organizational Research Methods,* 11: 79-95.

Wood, W., & Karten, S. J. 1986. Sex difference in interaction style as a product of perceived differences in competence. *Journal of Personality and Social Psychology,* 50: 341-347.