8

Performance Ratings: Then and Now

Frank Landy

Introduction

Over 30 years ago, Jim Farr and I finished one of the last large-scale manual literature reviews and narrative analyses of a central topic in industrial and organizational (I/O) psychology: performance rating (Landy & Farr, 1980). We considered hundreds of empirical and theoretical articles that appeared over the 30 years preceding our article. Since the appearance of that article, it has been cited over 500 times, so it clearly did and still does address a topic of interest to I/O psychologists. We drew a number of conclusions in that article. Some of the conclusions remain as true today as they were then and are hardly controversial. For example, we argued that performance rating was much more complicated than it might appear. We suggested a process model that included some of the complicating factors. Although that model can and has been improved, no one has suggested that rating is any *simpler* than we suggested. Even that preliminary model was likened by Jim Naylor to the plumbing in an old Scottish castle. We also suggested that cognitive operations of raters deserved serious consideration. Although this may have been a novel proposition for I/O psychologists, it was hardly earthshaking for the rest of the psychological research community. The cognitive revolution was well under way in most areas *other* than I/O psychology. Again, this proposition was embraced and, along with the work of Feldman (1981), could be seen as a valuable point of departure for later research.

Two other propositions, while apparently accepted at the time, have become more "controversial." The first was that a moratorium should be

declared on rating scale format. Since the beginning of this millenium, this proposition has been increasingly questioned. New technologies, new deconstructions of the performance domain, and new forms of work have led researchers to suggest that the moratorium should be lifted. As an example, Borman's introduction of computer-adaptive rating scales (CARSs; Borman et al., 2001; Schneider et al., 2003) shows great promise with respect to both the accuracy and the validity of ratings as well as the effectiveness of a new process for gathering information. I cannot speak for Jim Farr, but I am delighted with this line of research and willingly hereby officially lift the moratorium (as if it mattered). The second proposition that can be found in that article (Landy & Farr, 1980) dealt with the effect of demographic rater and rate characteristics, particularly race, gender, and age, on ratings. Simply put, Jim and I suggested that, from data available at the time, there was little clear evidence of bias on ratings based on demographic characteristics. For example, we said

- Rater and ratee demographic characteristics (ignoring possible moderator variables such as cognitive complexity or ratee familiarity) have little systematic effect on ratings.
- Rater and ratee demographic characteristics do not appear to interact to produce biased ratings.

These conclusions have been repeated many times and are often cited for the proposition that ratings are not biased against women, ethnic minorities, or older employees. This proposition of a lack of bias has become increasingly central to arguments of employment discrimination. Plaintiffs often suggest that performance ratings are unduly subjective and lend themselves to discriminatory decision making by managers and employers. They suggest underlying dynamics such as negative stereotypes or implicit attitudes.

In the 30 years since Jim Farr and I completed our literature review (Landy & Farr, 1980), substantial data have appeared on the topic of biased performance ratings. Better yet, many of the research designs have included both more realistic work settings and powerful analytic tools, such as meta-analysis, that provide a form of statistical control not widely available 30 years ago. Finally, the nature of the performance data is becoming more specific. For example, performance data are increasingly parsed into technical performance, Organizational Citizenship Behavior (OCB), counterproductive performance, and even adaptive and proactive performance. As a result, Jim Outtz asked me to revisit this arena and see if my conclusions would be the same today as they were 30 years ago. I think this was an excellent idea, and it forms the substance of my chapter.

The Literature Search and Review

The search for literature began with 1979 (the first year after the completion of the Landy and Farr article) and carried on to 2007. It was accomplished through PsychINFO (http://www.apa.org/psycinfo/), the American Psychological Association electronic database using the following key words: *performance appraisal, performance rating and race, gender, age, disability, bias*; *gender bias, age bias, race bias, disability bias*. Google Scholar (http://scholar.google.com/) was also searched using the same key words. Finally, a Social Sciences Citation Index (http://thomsonreuters.com/products_services/scientific/Social_Sciences_Citation_Index) search was completed using the Landy and Farr (1980) article as the search key. The search produced 230 empirical and theoretical articles. It is important to point out that these articles can be thought of as often "nested." Various meta-analyses often included the individual databases in their analytic scheme. Thus, one can consider the results of individual studies, the results of meta-analyses, or both. In my consideration, I do both. Nevertheless, it is instructive to know that of the 134 individual studies, 5 or 6.7% were also included in a meta-analysis. Naturally, a meta-analysis provides greater confidence of inference since it not only can control for statistical artifacts but also often tests possible moderator variables. That is not to say that the individual studies provide no unique insight on the phenomenon of interest (particularly when they include contextual variables that do not lend themselves to large-scale moderator subanalyses in meta-analyses), just that they are vulnerable to artifactual influences.

Table 8.1 presents the descriptive results of the literature search. As shown, some topics were of much greater interest to researchers than others. For example, gender variables produced the greatest number of empirical articles, while disability variables appear to be of less interest. Age and race fall somewhere in between. Similarly, meta-analyses have been completed on race and age, but less commonly gender and never disability. In part, this is an issue related to the coding of data in original studies. An analysis of gender, race, age, or disability can only be done if that variable is recorded at the individual rating level. In the subsequent

TABLE 8.1

Number of Studies 1997–2007

	Meta-analyses	Individual studies
Age	4	22
Gender	2	68
Race	4	40
Disability	0	14

sections of this chapter, I first consider meta-analyses and what one might conclude from them regarding bias in performance ratings. Next, I provide a sample of intriguing findings in individual studies. Finally, I address the larger issue of whether we know more now about potential rating bias than we did in 1980 and, of equal importance, whether there are forces at work that will limit the half-life of what we know today.

Performance Ratings and Race

Meta-Analyses

There have been three meta-analyses directly on point with respect to racial differences (commonly black–white differences) in performance ratings. Unfortunately, they each share a common flaw: The design was a between-subject design rather than a within-subject or repeated-measures design. The gold standard would be a meta-analysis of studies that used a repeated-measures design. This would mean that the stimulus objects (the employees) would be constant across raters. In the concrete, an example of this design would be black and white raters each rating the same black and white ratees. That way, we could be sure that any substantial Rater × Ratee race interaction was not due to simple cohort differences in which employees were rated. This design also allows us to identify main effects for both rater race and ratee race.

Thus, in considering the three meta-analyses on point, I found the following:

1. Kraiger and Ford, 1985: This meta-analysis dealt with the issues of performance ratings. All but 1 of the 88 studies included in the meta-analysis were between-subject designs and not able to address Rater × Ratee race effects that are critical for inferences about racial bias in ratings. In addition, although Kraiger and Ford examined some moderator variables (e.g., training, rating scale format, research setting), the small number of studies in which there were black raters (14 of 78) made it impossible to examine these moderators for black raters.
2. Ford, Kraiger, and Schectmann, 1986: This meta-analysis dealt with a comparison of race effects in performance ratings versus objective indices of performance. Again, the critical design flaw was the use of a between-subject design. The contaminating effect in ratings is the same as noted: There could have been real differences in performance rather than biased ratings. With

respect to objective indices, the contaminating effect was more subtle. If there was not a direct comparison on objective indices for blacks and whites at least *holding the same job title*, we are again unable to distinguish between true performance differences and differences biased by race of the subject.

3. Kraiger and Ford, 1990: In this meta-analysis, Kraiger and Ford used an indirect approach to study the question of possible bias on ratings: They examined the strength of the relationship between two different measures of job performance (objective indices and scores on job knowledge tests) and supervisory ratings for black and white employees. As was true in the earlier studies, the rating variable was a between-subject variable, and there was no way of ensuring that the differences in ratings between black and white raters were not the result of true performance differences. In other words, there was no way of estimating the Rater × Ratee race interaction effect—the gold standard for tests of bias.

Individual Studies

Even though meta-analyses of race differences in ratings are scarce (and all use a flawed design), there are some individual studies that have been done with more appropriate designs. I consider several of those studies and their findings. For none of the single studies I consider, whether for race or other demographic characteristics, will I consider studies that were done with student participants or with hypothetical employees. I have argued in other places that these studies are largely irrelevant for the purpose of drawing inferences about workplace decision making (Landy, 2005, 2008). They employ a stranger-to-stranger paradigm that suppresses the effect of individuating information. It is exactly this individuating information that characterizes the nature of workplace performance evaluations. I also do not deal with studies of assessment centers or employee development since they address other issues (and virtually all are between-subject designs).

There have been many individual studies of the main effects of race on performance evaluations. Most of those have used a between-subject design so are generally uninformative about the specific issue of narrowly construed "bias" in ratings. By narrowly construed, I mean the hypothesis that ethnic minorities receive unfairly low performance ratings. To test this proposition, it would be necessary to show that when exactly the same ratees are involved, there is a Ratee × Rater race interaction that works to the disadvantage of ethnic minorities. Note that even in these appropriate designs, it would be useful to know if majority ratings are unduly positive, minority ratings are unduly negative, or both. To date, even those studies that have uncovered a Rater × Ratee race effect cannot

answer that question. For all practical purposes, however, that is largely irrelevant to the extent to which minority employees may be deprived of scarce resources. If I am an ethnic minority, it hardly matters that whether I am unduly hammered in an appraisal or a majority member is unduly favored if I do not get a promotion or a pay raise.

To begin with individual studies that used only a between-subject design, there are some consistent findings, and they are consistent with the initial meta-analysis of Kraiger and Ford (1985). Generally, blacks received lower ratings than whites when race of rater was not crossed with race of ratee (Elvira & Zatick, 2002; Greenhaus & Parasuraman, 1993; Greenhaus, Parasuraman, & Wormley, 1990; Lefkowitz & Battista, 1995; Ostroff, Atwater, & Feinberg, 2004). The effect sizes were remarkably consistent at about 0.2 standard deviation (*SD*). Note that this consistency mirrors the consistency of differences in black–white differences in general mental ability (i.e., we consistently find differences of this magnitude), but it is a good deal lower in terms of effect size than the common finding of 1-*SD* difference in test scores. Another finding to emerge from these individual studies is that peer ratings showed larger race main effects than supervisory ratings (e.g., Pulakos, Oppler, White, & Borman, 1989). This might be argued as evidence that rater training (provided more often to supervisors than peers) and supervisory responsibility act as a check and balance against discrimination.

There have been several individual studies that have been completed using a repeated-measures design. Although this falls short of the power of a meta-analysis, most of these studies have been done with very large samples from both public sector and private sector employment covering many job titles and work contexts. As a result, we can place greater confidence in these findings than if they were from a small sample with one context and one job title. To be fair, it is not clear if there will ever be sufficient studies with appropriate information to do a meta-analysis to test for Rater × Ratee race effects. Such data are scarce in the field both because race of rater is not always coded and because the opportunities for having the same employees rated by both a minority and a majority supervisor are few and far between. Thus, the best we can hope for are large data sets.

As I described here, Kraiger and Ford (1985) were early meta-analytic investigators of possible race effects in ratings. After the appearance of their 1985 results, Pulakos, Oppler et al. (1989) conducted an analysis of military data from Project A. Pulakos et al. calculated point biserial correlations between race and ratings while controlling for the Rater × Ratee race interaction. They concluded that the "effect" of that interaction was much smaller (accounting for less than 1% of the rating variance) than the one found in the between-subject design of Kraiger and Ford.[1] A follow-on study (Oppler, Campbell, Pulakos, & Borman, 1992) examined the effect of objective indices of performance as a way of determining if any

main effects were due to favoring majority applicants, disfavoring minority applicants, or both. Generally, they found that minority–majority rating differences were mirrored by minority–majority objective criterion differences, rendering a bias interpretation less tenable. In addition, they concluded that supervisory ratings were more performance relevant than peer ratings. Finally, they raised an issue that had been largely ignored to that point: Maximal measures of performance (i.e., in this case objective performance indicators) demonstrated greater race effects than typical measures of performance (i.e., supervisory ratings).

Sackett and DuBois (1991) also questioned the Kraiger and Ford (1985) results from the perspective of experimental design. Sackett and DuBois correctly stated that a repeated-measures design was more helpful in teasing out any racial animus in ratings than a between-subject design. As a result, Sackett and DuBois analyzed both a large civilian and a large military database using, in part, a repeated-measures design. In addition, they reanalyzed the Kraiger and Ford data. Sackett and DuBois came to the following conclusions:

1. There was no statistical evidence of a Rater × Ratee race effect when employing a repeated-measures design; this was true both in the private sector and the military data. The earlier large-scale military analysis of Pulakos et al. (1989) had cautioned that the finding with military samples needed to be replicated with private sector data. Sackett and DuBois filled that private sector "hole."
2. When the Kraiger and Ford (1985) data were deconstructed to examine (a) the status of the raters (supervisors vs. peers), (b) the setting for the research (laboratory vs. field), and (c) the year in which the study was conducted (pre- vs. post-1970), the effect sizes became vanishingly small and in fact favored minority ratees slightly. Sackett and DuBois also controlled for peer versus supervisor ratings in their military data set.

Sackett and DuBois concluded that, when appropriately analyzed, rating data provided no evidence of a bias against minority ratees or evidence to suggest that raters provided higher ratings to ratees of their own race.

Waldman and Avolio (1991) analyzed a large data set derived from a U.S. Employment Service database covering many job titles and contexts. Although they could not do a true repeated-measures analysis, they did use a common criterion definition (supervisory ratings) as well as rater-ratee pairings by race. From their analysis, they concluded that the ratee differences reported by Kraiger and Ford (1985) were much larger than those found in their study. They attributed these smaller differences to controls they applied for ability, education, and job experience. In addition, they found much smaller Ratee × Rater race interactions (similar to

the findings of Pulakos et al., 1989 and Oppler et al., 1992). In essence, they concluded that when controls that might signal true score differences in the performance of whites and blacks were applied, differences in rating diminished to de minimis levels.

Rotundo and Sackett (1999) analyzed a large U.S. Employment Service performance-rating database with varied job titles and contexts. They had a slightly different focus in this study than the simple difference between black and white ratings. Instead, they sought to determine if bias ratings in a criterion could artificially influence validity coefficients involving cognitive ability tests. They were able to conduct both within- (repeated-measures) and between-subject analyses of these data. They found a very small combined effect size (0.07) between matched rater-ratee race data points and, based on hierarchical regression analyses, concluded that there was no evidence of bias on ratings; thus, it was unlikely that observed validity coefficients involving cognitive ability tests were artificially inflated by biased criterion scores.

Dewberry (2001) considered the potential presence of racial discrimination in the evaluation of written examination answers provided by white and black trainees in a legal education program. The wrinkle in this study was that the answers were graded in both a blind (to race) and nonblind condition by the same raters. Although Dewberry found small ratee effects (suggesting that black trainees performed more poorly than white trainees on the written examinations), there was no evidence of bias on the part of the raters (answer evaluators) since their black performance ratings were largely identical regardless of whether the ratings were blind or nonblind. It is axiomatic that blind ratings are unlikely to lend themselves to bias. Nevertheless, there were some common confounds in this study: It was not a repeated-measures design with respect to raters; plus, there was no study of rater race in either the blind or nonblind condition.

Stauffer and Buckley (2005) reanalyzed the Sackett and DuBois (1991) data set and concluded that Sackett and Du Bois were mistaken in their conclusions. Stauffer and Buckley showed that there were significant Rater × Ratee race effects that were both statistically (and they argued, practically) significant. Although there has been no formal response to this article by Sackett and DuBois (or others), there are some points than can be made in response.[2] As I have said here, when I consider the term *bias* in the context of performance ratings, I generally consider the narrow construction that addresses whether underrepresented groups (women, ethnic minorities, older employees, disabled employees) *suffer* as a result of biased ratings. More broadly and literally construed, bias can be seen as *any* differences between protected and nonprotected groups, even when the protected groups fare more favorably than the nonprotected groups. Stauffer and Buckley assumed the latter construction and argued that there are interaction effects but conceded that these effects do not

necessarily favor majority employees. Stauffer and Buckley noted that they do not know if any interaction is the result of more (unfair) favorable black ratings by black supervisors, less (unfair) favorable black ratings by white raters, or both. Stauffer and Buckley further argued that if the ratings disadvantage black employees, if the ratings determine who is considered "successful," and if the rating level necessary to be considered successful places more black employees below that level, then up to 12% of those employees could be inappropriately classified as "unsuccessful" with the possibility of palpable practical consequences. But, these hypotheticals do not deal with the complementary possibility that these black employees may actually be unfairly advantaged by higher-than-deserved ratings from black supervisors. Stauffer and Buckley concluded that there should be continued and vigorous research on the topic of race bias in performance ratings. As I argue in a concluding section of this chapter, there is every reason to agree with their plea, although not necessarily for the reasons they suggest.

Summary of Recent Literature on Race Bias in Performance Ratings

After reviewing recent meta-analyses on the possible bias of same-race raters on ratee performance evaluation, it appears that there are often significant differences between white and black mean ratings, to the disadvantage of the black ratees. Nevertheless, when we consider the results of within-subject designs and eliminate the effect of peer ratings and laboratory (student) ratings, any remaining Rater × Ratee race variance is small. Further, research that does not employ a within-subject design yet controls for attributes such as ability, education, and experience similarly points to small main effect black–white differences. Performance evaluations do not seem to be the type of egregiously subjective instruments that plaintiff lawyers allege. This does not mean that performance evaluations cannot be used as a pretext for unfair discrimination in a given instance. It does, however, suggest that there is nothing fundamental in the performance evaluation process that unleashes invidious negative race-based stereotypes

It would be valuable to see studies that examine Hispanic, Asian, and other ethnic minority subgroups to complete the picture.

Performance Ratings and Gender

Before addressing the research on performance ratings and gender, I make the following observation. Unlike race research, gender research

has centered on the dual issues of leadership and "male" and "female" jobs. These two movements are not independent. The reason for concentrating on leadership issues is likely the recognition that women are underrepresented in leadership and management positions—particularly senior leadership and management positions—and the assumption that this must be the result of invidious discrimination, most often laid at the feet of negative stereotypes of women as managers or leaders. Although this is certainly a noble and fruitful line of research, important issues related to performance ratings have been sidestepped, so we know a great deal less about the "observables" in the gender-related performance-rating research than we do about the possible consequences if there is a disadvantage that accrues to female workers in employment settings.

In addition, the gender performance evaluation research has largely been conducted in tightly controlled laboratory settings and to a lesser extent in field settings. Thus, students are asked to provide evaluations of hypothetical employees as seen in video or paper descriptions of their "work." As is true of demographic research to follow in this chapter (age and disability), there is little recognition in the gender research of the value of repeated-measures designs, which may tightly control at least the characteristics of the employee, or designs that control for experience, education, or abilities. This is unfortunate because we learn a lot from such designs, even though they are difficult to achieve in field settings. Nevertheless, it would appear that in the broad area of gender-related performance-rating research, it is probably easier to conduct a repeated-measures analysis than it is for race or certainly disability. Leaving aside the interest in leadership or senior management titles, there are plenty of female workers below those levels who could be (and often are) rated by both male and female supervisors. This was certainly true in the Pulakos et al. (1989) military study, and there is no reason to believe it could not be achieved in a counterpart private sector study. It is not my role to deconstruct *why* gender research has become centered on leadership issues, just to note that the empirical research is skewed toward those issues. And, one is left to wonder why race research did not follow a similar track since it is arguably true that a glass ceiling or a glass wall is just as pernicious for ethnic minorities as it is for women.

Meta-Analyses

Bowen, Swim, and Jacobs published a meta-analysis of gender-based performance-rating research in 2000. In that publication, they identified previous meta-analyses and reviewed them in some detail. I do not reprise the Chieh-Chen et al. review of other meta-analyses, but I summarize some of the salient points that they made regarding those other meta-

analyses since it clearly sets the limits for what I cover in my review of both meta-analyses and individual studies.

1. Meta-analyses conducted by Swim, Borgida, Muruyama, and Myers (1989); Eagly, Makhijani, and Klonsky (1992); Olian, Schwab, and Hammerfield (1988); and Davison and Burke (2000) analyzed only laboratory studies.
2. With the exception of the Olian et al. (1988) meta-analysis, all considered the sex stereotype of the job in question (in the studies included in the meta-analysis) as a control variable, a manipulated variable, or a moderator variable.
3. With the exception of the Olian et al. (1988) meta-analysis, all found slight advantages to women in ratings.
4. In several of the meta-analyses (Davison & Burke, 2000; Eagly et al., 1992; Swim et al., 1989), women fared better in masculine-stereotyped jobs and when rated by males. This is puzzling from the perspective of negative female stereotypes since females should have been rated more harshly in male-stereotyped jobs and by male raters if the proposed stereotypes were operating as proposed.

Eagly, Karau, and Makhijani published a meta-analysis of mixed field and laboratory studies of leader effectiveness in 1995. The dependent variables that were examined were the rated satisfaction with or the performance of leaders. Necessarily, the performance dimensions dealt with leader-related behaviors rather than broader issues related to technical performance, citizenship behavior (at least directly), counterproductive behavior, or adaptive behavior (Landy & Conte, 2004, 2007). Laboratory studies included both Goldberg paradigm designs (presentation of resumes) or ad hoc group interactions. Since both of these scenarios involve stranger-to-stranger paradigms, I concentrate on the field studies rather than the laboratory studies because the field studies more directly address the point of this chapter. The studies in the meta-analysis included 10 military samples. Since the Pulakos et al. (1989) study described in the race section of this section and elsewhere in this section was not included in the meta-analysis, I presume it was because Pulakos et al. did not address leadership directly. Nevertheless, as we shall see, Pulakos et al. found little gender effect in performance ratings.

Of the 74 studies that were classified as "organization," 22 were classified as "business," 21 as "educational," 7 as governmental or social service, 10 as military, and 14 as miscellaneous. For the sake of this review, I accept the organizational category as the appropriate level (although I comment, as do Eagly et al. (1995), on the apparent uniqueness of the

military sample). The results of the meta-analysis with respect to organizational studies showed little of the effect that had been found in the meta-analysis of race by Kraiger and Ford (1985), at least with respect to leadership ratings. There were some modest effect sizes (e.g., $d = 0.05$) suggesting lower ratings for female leaders, but when these were further disaggregated into organizational versus other, there were small advantages for women leaders in ratings in all categories (ranging from $d = 0.05$ to $d = 0.15$), except military settings, for which there was a pronounced advantage for male leaders ($d = 0.42$).

The conclusions to be drawn from the Eagly et al. (1995) meta-analysis are best stated by the authors themselves:

> When all of the studies in our sample were aggregated, female and male leaders did not differ in effectiveness. [This] suggests that despite barriers and possible handicaps in functioning as leaders, the women who actually serve as leaders and managers are in general succeeding as well as their male counterparts. (p. 137)

So, with respect to the purposes of this chapter, we may conclude that in the limited (but important) world of leader behavior, women seem not to be disadvantaged in performance ratings. Although there were military data to suggest that women leaders are rated lower than male counterparts, I reserve discussion of this point for a reconsideration of the Pulakos et al. (1989) study of female soldiers.

Chieh-Chen et al. (2000) conducted a meta-analysis of the broader topic of gender effects on performance ratings in field studies. They did not further confine the study to leadership ratings. Like other meta-analyses (of race and gender), there were no analyses that compared between-subject designs with repeated-measures designs. This means that it was not possible to control for ratee characteristics by matching rater gender with ratee gender directly. Nevertheless, the authors did consider controls for organizational level, experience, and education, although not for ability as had been done in some earlier research on race. Chieh-Chen et al. specifically targeted field studies for their meta-analysis. Using various selection rules, 32 study samples were analyzed in the meta-analysis. As had been found by Eagly et al. (1995), there were only small effect sizes for ratee gender, and they slightly favored female ratees. Further, no significant effect sizes were discovered for masculine- versus feminine-typed jobs or for the relative proportion of male ratees in a work group (a significant effect for group composition is often interpreted as evidence for "tokenism" in performance ratings). Rater training and individuating information both tended to decrease any bias in the ratings (although the effect of training appeared to *decrease* pro-female bias). Although the stereotypicality of the job (male or female) influenced ratings, it appeared to simply *decrease* the

female rate advantage. This has been an argument of many proponents of stereotyping processes. They argue that perhaps women *should have an advantage* when certain performance domains (e.g., citizenship, interpersonal skills) are considered. This simply further amplifies the need for repeated-measures designs in deconstructing the effect of gender on performance ratings.

Chieh-Chen et al. concluded that "there is little systematic evidence of overall gender bias in performance evaluations in actual work settings" (p. 2205). Nevertheless, they did caution that when the stereotypicality of the performance measure is taken into account and the gender composition of the raters is taken into account, there does seem to be some evidence of lower ratings for women. The absence of the repeated-measures design element renders this caution less dramatic than it might seem. The authors did, however, reasonably ask that if we take the results of performance appraisals at face value, then we may question why there is not a similar advantage to women when applying for promotions in an organization, but this is the meat for another time and another chapter.

Individual Studies

In the section on race, I described the large-scale study conducted by Pulakos et al. (1989) as part of the Project A effort. I do not repeat that description. In addition to race, Pulakos et al. studied gender, using both a between- and a within-subject (repeated measure) design. Unlike the Eagly et al. (1995) meta-analysis, the Pulakos et al. analysis did not address leadership issues but instead analyzed ratings of technical skill and job effort, personal discipline, and military bearing. The technical skill and job effort dimension did include a consideration of "demonstrating leadership and support toward peers." Unlike the race analyses performed, which included large and equal numbers of blacks and whites, the gender analyses included many fewer females as raters and ratees as compared to males. In the repeated-measures analysis of supervisory ratings, there was neither a significant main effect for gender nor a significant Rater × Ratee gender interaction effect.

Although the repeated-measures design remains the gold standard for examining demographic effects on performance ratings, it is useful to consider studies of main effects as well. When considering field studies, and eliminating those studies included in the Chieh-Chen et al. (2000) and Eagly et al. (1995) meta-analyses, the results of single studies are interesting. These studies did not appear in the meta-analyses either because they appeared after the meta-analysis was completed or because they did not have the accompanying information sought by the meta-analysts. Nevertheless, they largely confirmed the results of those meta-analyses.

One subset of these studies showed no main effect for ratee gender or no Rater × Ratee gender interactions. These include work by Lefkowitz and Battista (1995); Shore and Thornton (1986); Shore, Tashchian, and Adams (1997); and Sinangil and Ones (2003). A larger subset of these individual studies reported main effects favoring female ratees. These include studies by Furnham and Stringfield, 2001; Lewis, 1997; Ostroff et al., 2004; and Shore, 1992. In a related study, Guetal, Luciano, and Michaels (1995) examined the extent to which pregnancy might uniquely stigmatize and unfavorably influence female performance ratings. They discovered that performance ratings actually increased during pregnancy when compared to prepregnancy ratings for the same employees and when compared to control groups of nonpregnant women who held matched job titles and who were evaluated at the same time as the pregnant women.

There were two other individual studies that were not so simple to categorize. Sackett, DuBois, and Wiggins Noe (1991) examined the possible role of tokenism in ratings of both blacks and women. The results were nuanced. No tokenism effects were found for blacks; that is, the racial composition of the work group played no role in ratings. But, for gender, when women made up 20% or fewer of the work group in question, they were rated about 0.5 *SD* lower than men. In contrast, when women represented more than 50% of the work group, they were actually rated *higher* than their male counterparts. The authors explained the nuanced effects of token status by invoking the possibility that certain "jobs" (i.e., those with fewer than 20% female incumbents) may be more "masculine" in type than the jobs in which females predominate. If that is the case, tokenism is less likely the explanation than gender typing of jobs, as has been suggested by Eagly, Heilman, and others. Further, Sackett et al., using regression analysis, found that when controls were put in place for education, ability, and firm experience, gender composition accounted for an additional 4% of the variation in performance ratings. So, we are left with an intriguing individual study that raised more questions than it answered.

A second study that bears attention was conducted by Lyness and Heilman (2006). In this study, the variable of interest was the fit between gender and the line versus staff nature of a position. Line positions were considered stereotypically male, while staff positions were considered stereotypically female. A follow-on analysis examined the promotional history of women in line and staff positions and their performance ratings. The results were interesting. According to the authors, "Women in managerial line jobs received lower ratings than women in managerial staff jobs or men in either managerial line or staff jobs but promoted women had received higher performance ratings than promoted men." The research design included neither a repeated-measures aspect, which would permit a Rater × Ratee gender interaction analysis, nor controls for ability or experience (although

they did control for organizational tenure and age). Thus, as the authors correctly noted, one cannot rule out true performance differences between the female line managers and the comparator groups. The most appropriate summary of the study results might be that there are some special circumstances (upper level, female, line managers) in which bias may occur even though in aggregate there is no evidence of gender bias in ratings.

Summary of Recent Literature on Gender Bias in Performance Ratings

If, as I suggested, we narrowly construe the notion of bias to mean that women are disfavored in performance ratings, the available evidence suggests that this is not true, and in fact women are more likely to receive higher ratings than men, all other things being equal. This is different from race, for which we could conclude that there were no *disadvantages* to ethnic minority status in terms of ratings. Nevertheless, there are three equivocal possibilities for exception. I use the word *equivocal* because there are several potential explanations other than gender for the results. The three possibilities are (a) upper-level female managers are viewed differently from lower-level female managers; (b) female line managers are viewed differently from female staff managers; and (c) when women make up less than 20% of a work group, they may occupy a special stigmatized position.

But in general, my conclusion based on 30 years of research on the question is similar to the conclusion that Jim Farr and I drew in 1980: The evidence of any systemic discrimination in the ratings of working women is scarce and possibly localized.

Performance Ratings and Age

Unlike gender or race, age is (unfortunately) not an immutable demographic characteristic. Those who were young will eventually become old (barring untimely death). This means that research designs have the additional option of longitudinal and time-lagged cohort analyses that are not available in the study of race or gender. Although longitudinal analyses of ratings for gender and race could be completed, they would inevitably be confounded by age. One does not become more female or more Hispanic over time, but one does become older. As was the case in the study of race and gender, repeated-measures designs are just as valuable and just as informative for age as they are for other demographic characteristics. Unfortunately, age researchers seem as constrained to conventional between-subject designs as race and gender researchers.

Meta-Analyses

There have been four meta-analyses of the relationship between age and performance ratings since the late 1980s. I review these studies next.

Waldman and Avolio (1986) analyzed 40 samples of relationship of age to performance data. Using correlational analysis, the authors reported that age accounts for approximately 2% of the rating variance, and that older workers generally receive lower ratings. When positions classified as professional are distinguished from those classified as nonprofessional, the percentage of variance in professional ratings associated with age drops to near zero, while the percentage of variance associated with age for nonprofessionals remains at approximately 2%. The authors were not able to rule out the possibility that even these modest associations were not true performance differences.

In 1989, McEvoy and Cascio conducted a meta-analysis of 96 independent studies (including some of the same studies included in the Waldman & Avolio, 1986, analysis described in the preceding paragraph). McEvoy and Casio found a very small correlation between age and performance, accounting for less than 1% of the rating variance. Unlike the Waldman and Avolio results, McEvoy and Cascio found no effect for professional versus nonprofessional status.

The third meta-analysis was conducted by Finkelstein, Burke, and Raju (1995) and, like the meta-analysis of gender by Davison and Burke (2000), only considered laboratory experiments rather than field studies. As a result, I do not review this meta-analysis in any detail. I note that when individuating information was provided to experimental subjects, bias was drastically reduced, a common finding in stereotyping research.

The fourth meta-analysis was conducted by Gordon and Arvey in 2004. An analysis of 52 samples (including both laboratory and field studies, and including many of the studies that appeared in earlier meta-analyses) revealed an overall effect size of $d = 0.11$. This is considered small. Using publication date as a moderator variable, they found that there was less evidence of age bias in more recent than in more distant studies. In a comparison of laboratory versus field studies, the researchers found considerably more evidence of bias in laboratory studies using student raters. Gordon and Arvey concluded that when raters are supervisors, when there is ample information about the ratees, and when the data were collected recently, there was little evidence of age bias in ratings.

Individual Studies

I was able to identify only two individual studies that were either not laboratory studies or not included in the meta-analyses described. The first (Vecchio, 1993) examined the situation of (teacher) subordinates who were

older than their (principal) supervisors. He found that there was a nonsignificant association between teacher age and principal's evaluations of the teacher performance. More specifically, in examining the scenario of older subordinates and younger supervisors, he found no evidence of age bias against those older subordinates. In the second study, a related study of age similarity/dissimilarity between supervisors and subordinates, Ferris, Judge, Chachere, and Liden (1991) found no evidence of same-age bias; instead, ratings were more favorable when there was a *dissimilarity* in the age of supervisors and subordinates.

Summary of Recent Literature on Age Bias in Performance Ratings

As we have seen in race and gender analyses, there is no evidence to suggest that older workers receive significantly lower performance ratings than younger workers. In fact, when these older workers have younger supervisors, it appears that these older workers may actually receive higher ratings. The bad news is that the research designs used to study age influences on ratings are either inappropriate (i.e., cross sectional rather than longitudinal or cohort based) or lack control (of experience, ability, education, etc.). The good news is that there is no age-related variance to partial out of associations.

Performance Ratings and Disability

Disability is a relatively new focus of work-related research. The Americans With Disabilities Act (ADA) was passed in 1990, thus it is the "new" statute on the block compared those for to race, gender, and age, which have longer-standing statutory protections. As a result, the research database is embarrassingly sparse. There have been no meta-analyses and few empirical field studies of ratings of the disabled. Without research, one must concede that disability has a special status in America and many nations. Assumptions regarding capabilities seem so likely that the ADA even incorporates a protection and a claim of action for the "perception" of a disability by an employer. Thus, I think that we might assume a priori that disabled workers will receive lower performance evaluations than their more able counterparts. But as scientists, we deal with empirical confirmations of hypotheses, not impassioned speculation. Most data related to disability and performance judgments come from laboratory experiments asking students to assume the role of an employer. These studies hold no value for the present review. Other publications are largely descriptive, cataloguing the indignities suffered

by disabled workers. There are a few studies that surveyed employers about their concerns regarding the hiring of disabled workers. As one might expect, potential and actual employers were concerned about ineffective performance (e.g., Johnson, Greenwood, & Shriner, 1988; Smith, Webber, Graffam, & Wilson, 2004; Tse, 1994). Nevertheless, sad to say, there are simply no data available to address the issues of possible bias in performance ratings of disabled workers.

General Summary and Conclusions

Thirty years ago, Jim Farr and I suggested that with the data available to us, we saw no substantial evidence of bias in performance ratings related to demographic characteristics of ratees or raters. Based on a review of hundreds of studies, meta-analyses, regression analyses controlling variously for ability, education, experience, and organizational level and job type, and designs incorporating repeated measures as a control for true performance levels, I find no reason to change that conclusion.

Nevertheless, this is not a call for a moratorium on anything—I have learned that lesson at least. If for no other reason, the parsing of the performance domain into technical, citizenship, counterproductive, and adaptable facets of performance signals a need for new analyses or reanalyses of these noneffects (including fresh meta-analyses of old data that permit such a parsing) to see if they remain noneffects. Further, the changing nature of work (larger spans of control, more team-oriented work, etc.) suggests additional moderator variables to examine. I encourage such research. Further, the intriguing findings related to gender-job stereotypes and work group gender composition require continued investigation. But, I would still argue that these more complex analyses should be accompanied by more basic analyses of Ratee × Rater gender interaction effects using repeated measures and control variables in the conduct of the research.

This fresh view of the performance-rating literature provides little foundation for a broad claim in the litigation context that performance ratings are inherently unfair to protected groups. That is not to say that, in a given instance, a performance rating was not used as a pretext for invidious discrimination. But, it is to say that there is no undue cause for alarm when performance ratings are assigned to protected groups—at least those defined by race, gender, or age.

Acknowledgments

I gratefully acknowledge the assistance of Jacob Seybert, Barbara Nett, and Kylie Harper in article search and production. Kylie Harper also assisted in editing drafts of this chapter. Kevin Murphy, Paul Sackett, Rick Jacobs, Jeff Conte, and Jim Farr provided thoughtful reviews of a draft of the chapter.

Notes

1. In a conversation with Elaine Pulakos, she expressed a reservation about her work and, more importantly, about other "field or operational" studies. I have made the point that an important distinction is between laboratory and field studies in part because laboratory studies have little of the accountability factor that field operational studies have. She was concerned that many field studies are carried out in the context of research studies, and participants are promised anonymity and that their ratings will not "count" even though they represent real and intact supervisor-subordinate dyads. Thus, she suggested, that the actual Rater × Ratee interaction might be higher than she estimated in her earlier studies and in later studies (e.g., Sackett & Dubois, 1991). This suggests that a new meta-analysis moderator should be examined that distinguishes between the truly operational (these "count") and the "research operational" (these do not count) to see if estimates of psychometric bias area are affected by this contextual issue.
2. I acknowledge the thoughts of Paul Sackett on the Stauffer and Buckley article (2005) in a recent personal communication to me.

References

Americans With Disabilities Act of 1990, 42 U.S.C.A. § 12101 et seq. (West 1993).

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86*, 965–973.

Bowen, 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*(10), 2194–2215.

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

Dewberry, C. (2001). Performance disparities between whites and ethnic minorities: Real difference or assessment bias? *Journal of Occupational and Organization Psychology, 74*, 659–673.

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

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

Elvira, M. M., & Zatick, C. D. (2002). Who's displaced first? The role of race in layoff decisions. *Industrial Relations, 41*, 329–361.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive process in performance appraisal. *Journal of Applied Psychology, 66*, 127–148.

Ferris, G. R., Judge, T. A., Chachere, J. G., & Liden, R. C. (1991). The age context of performance-evaluation decisions. *Psychology and Aging, 6*, 616–622.

Finkelstein, L. M., Burke, M. J., & Raju, N. S. (1995). Age discrimination in simulated employment contexts: An integrative analysis. *Journal of Applied Psychology, 80*, 652–663.

Ford, J. K., Kraiger, K., & Schechtmann, 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.

Furnham, A., & Stringfield, P. (2001). Gender differences in rating reports: Female managers are harsher raters, particularly of males. *Journal of Managerial Psychology, 16*, 281–288.

Gordon, R. A., & Arvey, R. D. (2004). Age bias in laboratory field settings: A meta-analytic investigation. *Journal of Applied Social Psychology, 34*, 468–492.

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

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

Guetal, H. G., Luciano, J., & Michaels, C. A. (1995). Pregnancy in the workplace: Does pregnancy affect performance appraisal ratings? *Journal of Business and Psychology, 10*, 155–167.

Johnson, V. A., Greenwood, R., & Shriner, K. (1988). Work performance and work personality: Employer concerns about workers with disabilities. *Rehabilitation Counseling Bulletin, 32*, 40–57.

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

Kraiger, K., & Ford, J. K. (1990). The relation of job knowledge, job performance and supervisory ratings as a function of ratee race. *Human Performance, 3*, 269–279.

Landy, F. J. (2005). *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives*. San Francisco: Jossey-Bass.

Landy, F. J. (2008). The tenuous bridge between research and reality: The importance of research design in inferences regarding work behavior. In E. Borgida and S. T. Fiske (Eds.), *Beyond common sense: Psychological science in the courtroom* (pp. 341–352). Malden, MA: Blackwell.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107.

Lefkowitz, J., & Battista, M. (1995). Potential sources of criterion bias in supervisor ratings used for test validation. *Journal of Business and Psychology, 9*, 389–414.

Lewis, G. B. (1997). Race, sex, and performance ratings in the federal service. *Public Administration Review, 57*, 479–489.

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*, 777–785.

McEvoy, G. M., & Cascio, W. F. (1989). Cumulative evidence of the relationship between employee age and job performance. *Journal of Applied Psychology, 74*, 11–17.

Olian, J. D., Schwab, D. P., & Hammerfield, 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.

Oppler, S. H., Campbell, J. P., Pulakos E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. *Journal of Applied Psychology, 77*, 201–217.

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

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

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

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

Sackett, P. R., Dubois, C. L. Z., & Wiggins Noe, A. (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.

Schneider, R. J., Goff, M., Anderson, S., & Borman, W. C. (2003). Computerized adaptive rating scales for measuring managerial performance. *International Journal of Selection and Assessment, 11*, 2–3, 237–246.

Shore, T. H. (1992). Subtle gender bias in the assessment of managerial potential. *Sex Roles, 27*(9–10), 499–515.

Shore, T. H., Tashchian, A., & Adams, J. S. (1997). The role of gender in a developmental assessment center. *Journal of Social Behavior and Personality, 12*, 191–203.

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

Sinangil, H. K., & Ones, D. S. (2003). Gender differences in expatriate job performance. *Applied Psychology: An International Review, 52*, 461–475.

Smith, K., Webber, L., Graffam, J., & Wilson, C. (2004). Employer satisfaction with employees with a disability: Comparisons with other employees. *Journal of Vocational Rehabilitation, 21,* 61–69.

Stauffer, J. M., & Buckley, M. R. (2005). The existence and nature of racial bias in supervisory ratings. *Journal of Applied Psychology, 90*, 586–591.

Swim, J. K., Borgida, E., Muruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin, 105*, 409–429.

Tse, J. (1994). Employers' expectations and evaluation of the job performance of employees with intellectual disability. *Australia and New Zealand Journal of Developmental Disabilities, 19*, 139–147.

Vecchio, R. P. (1993). The impact of differences in subordinate and supervisor age on attitudes and performance. *Psychology and Aging, 8*, 112–119.

Waldman, D. A., & Avolio, B. J. (1986). A meta-analysis of age differences in job performance. *Journal of Applied Psychology, 71*, 33–38.

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