

The Language of Performance Evaluations: Gender-Based Shifts in Content and Consistency of Judgment

Monica Biernat, M. J. Tocci and Joan C. Williams
Social Psychological and Personality Science published online 18 July 2011
DOI: 10.1177/1948550611415693

The online version of this article can be found at:
<http://spp.sagepub.com/content/early/2011/07/15/1948550611415693>

Published by:



<http://www.sagepublications.com>

On behalf of:

Society for Personality and Social Psychology



Association for Research in Personality

ASSOCIATION FOR
RESEARCH IN PERSONALITY

European Association of Social Psychology



European Association
of Social Psychology

Society of Experimental and Social Psychology



Additional services and information for *Social Psychological and Personality Science* can be found at:

Email Alerts: <http://spp.sagepub.com/cgi/alerts>

Subscriptions: <http://spp.sagepub.com/subscriptions>


Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

>> [Proof](#) - Jul 18, 2011

[What is This?](#)

The Language of Performance Evaluations: Gender-Based Shifts in Content and Consistency of Judgment

Social Psychological and
Personality Science
000(00) 1-7
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1948550611415693
http://spps.sagepub.com


Monica Biernat,¹ M. J. Tocci² and Joan C. Williams³

Abstract

Performance evaluations of male and female junior attorneys in a Wall Street law firm were analyzed. Male supervisors judged male attorneys more favorably than female attorneys on numerical ratings that mattered for promotion but offered narrative comments that showed either no sex effects or greater favorability toward women. Judgments of male attorneys were more consistent overall than they were for female attorneys, and predictors of numerical ratings differed by sex: Narrative ratings of technical competence mattered more for men than women, and narrative ratings of interpersonal warmth mattered more for women than men. Open-ended use of positive performance words—the only outcome that favored women—did not translate into positive numerical ratings for women. The data suggest subtle patterns of gender bias, in which women were harmed by not meeting gendered expectations of interpersonal warmth but were less benefited than men by meeting masculine standards of high technical competence.

Keywords

stereotypes, social judgment, sexism, prejudice/stereotyping, language

Modern forms of gender bias often take subtle rather than blatant forms. Rather than across-the-board discrimination against women in the workplace, backlash may occur against women who are perceived as too agentic (Heilman, 2001; Rudman, 1998; Rudman & Fairchild, 2004) or may take the form of overweighting of attributes that men possess (Uhlmann & Cohen, 2005) and paying greater attention to negative workplace behaviors of women than men (Biernat, Fuegen, & Kobrynowicz, 2010).

We suggest another form of gender bias in the workplace: Differential patterns of employee sex effects on numerical and narrative evaluations. Numerical ratings are among the most common forms of performance evaluation (Murphy & Cleveland, 1995), and in some work settings, these are supplemented with narrative commentary. Our question is whether and how these forms of evaluation relate to each other: Do positive narrative comments map onto positive numerical evaluations? Does employee sex moderate these patterns? We address these questions by examining the evaluations received by male and female junior attorneys working at a Wall Street law firm from their male supervisors. Given the “masculine” nature of finance law, we expect that male attorneys will receive more favorable performance evaluations—the consequential numbers that matter for promotion—than female attorneys (see Landy & Farr, 1980). Both Heilman’s (1983) “lack of fit” perspective and Eagly and Karau’s (2002) role congruity theory suggests that gender stereotyping of women will be most evident

in masculine domains: “negative expectations resulting from perceptions of lack of fit detrimentally affect how women are regarded and how their work is evaluated when they are in traditionally male jobs” (Lyness & Heilman, 2006, p. 777).

An additional question is whether this pattern holds across other types of evaluation and feedback offered to employees. The literature on performance evaluations suggests that correspondence between types of evaluation may be modest. For example, a meta-analysis by Heneman (1986) found an average correlation of .27 between subjective and objective (e.g., sales volume, output) ratings of employee performance; a more recent meta-analysis reported a slightly stronger effect, $r = .39$ (Bommer, Johnson, Rich, Podsakof, & MacKenzie, 1995). Other studies have shown that correlations between subjective ratings and other indicators of performance may be moderated by ethnicity or sex of the ratee and/or rater (e.g., Castilla, 2008; Kraiger & Ford, 1990).

¹ University of Kansas, Lawrence, KS, USA

² Fulcrum Advisors, Pittsburgh, PA, USA

³ University of California Hastings College of the Law, San Francisco, CA, USA

Corresponding Author:

Monica Biernat, University of Kansas, 1415 Jayhawk Blvd, Lawrence, KS 66045, USA

Email: biernat@ku.edu

In the present study, we compare numerical ratings and the open-ended narratives that accompany them. Typically, the purpose of narrative comments is for supervisors to support their bottom-line ratings (see Murphy & Cleveland, 1995), but as Wilson (2010) notes, supervisors may use these comments “to justify or mask their biases” (p. 1909). In her study of ethnicity effects in performance evaluations of bank staff in the United Kingdom, Black workers received lower numerical ratings than White staff members but narrative comments were overwhelmingly positive: “Supervisors systematically gave lower ratings to Black staff relative to White staff that they did not explain in their written summaries” (p. 1925).

Wilson (2010) suggests that these patterns occur because numerical ratings are made automatically, perhaps with reference to group stereotypes, whereas the writing of narrative comments requires greater cognitive effort and offers the opportunity for controlled or calculated responses. For example, one might use the narrative as an opportunity to “soften the blow” of a harsh evaluation. This might reflect a kind of paternalism or benevolent sexism in the case of gender (Glick & Fiske, 1996).

Additionally, narratives may allow evaluators to be “slippery” in their use of language; to apply similar words to mean different things, depending on the sex or race of the person being described. Members of negatively stereotyped groups may be compared to lower standards and therefore fare better in these comments (e.g., “good” for a woman may not be as good as “good” for a man; Biernat, 2003). But the positivity prompted by low standards may not translate into enhanced access to opportunities or resources. For example, women softball players were praised more than male players for hitting a single but were less likely to be placed in valued fielding or batting positions (Biernat & Vescio, 2002); female subordinates were praised more than male subordinates but were assigned to less valued positions on a work team (Vescio, Gervais, Snyder, & Hoover, 2005). In the Wall Street firm we consider in the present study, we expected that the pro-male bias in numerical ratings would be reduced or reversed in the narrative commentary.

Additionally, we expected lesser correspondence between numerical ratings and positive narrative comments for female than male attorneys. Others have described a pattern of performance-reward bias, which occurs when “women and minority employees receive different rewards for the same merit scores as White men” (Castilla, 2008, p. 1520). In Castilla’s study of performance appraisal at a large service organization, performance ratings predicted salary increases more strongly for White men than for women and minorities. In the study reported here, we do not assess the evaluation–salary link; rather, we examine whether positivity in narrative commentary better predicts the numerical ratings of male than female attorneys.

We also addressed the association between numerical ratings and specific attorney characteristics mentioned in narrative comments. A number of studies suggest that competence and warmth are key dimensions of interpersonal evaluation;

these factors seem to capture much of the “space” into which our perceptions of others fall (for reviews, see Cuddy, Fiske, & Glick, 2007; Fiske, Cuddy, Glick, & Xu, 2002; Judd, James-Hawkins, Yzerbyt, & Kashima, 2005). Gender stereotypes generally map onto this distinction, with women characterized as particularly high in warmth and men as high in competence (e.g., see Kite, Deaux, & Haines, 2007; Rudman & Glick, 2008; Spence & Buckner, 2000). Therefore, we expected that numerical ratings would be more closely tied to narrative mentions of warmth for female than male attorneys and more closely tied to narrative mentions of technical competence for male than female attorneys (e.g., see Sidanius & Crane, 1989).

Our study focuses on one law firm that invited a consultant to evaluate its evaluation procedures, and one round of evaluations received by junior attorneys. We therefore offer a “snapshot” of how attorney sex matters for performance evaluations in this firm, but we admit that the generalizability of our results to other firms and settings is unknown. Nonetheless, this context provides an important ground for examining gender bias using consequential workplace evaluations of real employees observed over a period of real time and high contact (see Dipboye, 1985).

Method

Actual performance evaluations of all 268 junior attorneys working in a single Wall Street law firm were obtained by the second author (M.J.T.) who was hired by the firm as a consultant to evaluate its annual review procedures. The data were collected in 2006 as part of the regular performance review process. The firm can be characterized as a typical Wall Street firm and perhaps typical of mid-size to large New York law firms. As part of the consultancy contract, clients agree to allow for publication of findings as long as anonymity is maintained.

Each junior attorney had been evaluated by senior lawyers in the firm (referred to throughout as evaluators) with whom they had had some work contact. Each evaluator offered numerical judgments of and open-ended narrative comments about the target attorney. These individual judgments were not made available to the junior attorneys but did form the basis of a summary evaluation each attorney received at the end of the review period. The junior attorneys were predominantly White (75.4%) or Asian (13.4%), and our analyses focused on this subgroup of 234 (84 female and 150 male) attorneys. The vast majority of the evaluators were men, and indeed, we focus here only on evaluations offered by these senior supervising male attorneys (mean [M] number of evaluators = 3.59, standard deviation [SD] = 2.31, range = 0–22).¹

Numerical judgments were made on 26 dimensions, using 1–5 response scales where 1 = *seriously below expectations*, 2 = *below expectations*, 3 = *meets expectations*, 4 = *exceeds expectations*, and 5 = *outstanding*. The 26 judgments were divided by the firm into 7 categories: technical excellence, effective lawyering, delivery of client services, teamwork and leadership, delegation and training of subordinates, attitude and personality, and dedication to the firm’s mission. The items

Table 1. Correlations Among Evaluation Components, by Attorney Sex

	Numerical Rating	Narrative Technical	Narrative Warmth	Positive Performance	Partner
Numerical rating	–	.63*	.65*	–.06	.21*
Narrative technical competence	.72*	–	.55*	.13	.16
Narrative warmth	.63*	.70*	–	.02	.13
Positive performance words	.17*	.30*	.22*	–	.14
Partner likelihood	.51*	.33*	.39*	.04	–

Note: Correlations above diagonal and boldfaced values = female attorneys ($N_s = 83-86$); below diagonal = male attorneys ($N_s = 147-152$). * $p < .05$.

were highly intercorrelated, r_s from .76–.93, strongly suggestive of a halo effect rather than nuanced discrimination among performance factors. To reduce the data, we computed the average numerical rating across all 26 dimensions ($\alpha = .98$) and refer to this index as the “numerical rating.” This is the most consequential of the evaluations attorneys receive as numerical ratings matter for outcomes such as raises and promotion: Common knowledge in the firm is that only attorneys who receive mostly “5s” are headed toward partnership.

In addition to the numerical ratings, evaluators offered open-ended narrative comments about the attorneys. These were analyzed in two ways. First, the narratives were submitted to a text analysis software system (Linguistic Inquiry Word Count; Pennebaker, Francis, & Booth, 2001), which performs counts of many types of word categories. Our interest was in the positivity of these comments, so we focused on the frequency of positive performance words (including excellent, good, awesome, terrific, stellar, wonderful, etc.). There was no difference in the total number of words written about female and male attorneys, $t < 1$ ($M_{\text{male}} = 386$, $SD = 285$; and $M_{\text{female}} = 394$, $SD = 304$).

Second, we trained independent coders to rate the open-ended descriptions, with gendered pronouns and any other references to gender removed. Two law students rated each narrative comment for technical competence and interpersonal warmth (using 1–5 [*very negative*–*very positive*] rating scales). Technical competence included references to analysis skills, drafting, legal judgment, efficiency, and productivity; interpersonal warmth included references to friendliness, warmth, attentiveness, responsiveness, and good communication/relationships with coworkers. Two undergraduate student coders also rated each block of text on overall warmth/friendliness and overall technical/drafting skills using the same scales. They also indicated whether the evaluator made any mention of partnership, using a 3-category system (–1 = *negative mention*, e.g., “not partner material;” 0 = *no mention*, 1 = *positive mention*, e.g., “on the right track to partnership”).

We averaged across the four coders’ judgments of technical competence/drafting skill to compute an overall “narrative rating of technical competence” index (Krippendorff’s $\alpha = .63$) and across the four coders’ warmth/friendliness ratings to produce a “narrative rating of interpersonal warmth” index (Krippendorff’s $\alpha = .61$). The two undergraduate coders’ partnership likelihood judgments were combined to produce a “partner likelihood” index (Krippendorff’s $\alpha = .81$).²

Results

Evaluations of Male and Female Attorneys

As predicted, male attorneys ($M = 3.90$, $SD = 0.54$) were judged more favorably than female attorneys ($M = 3.76$, $SD = 0.46$) on the numerical rating index, $t(232) = 1.98$, $p < .05$, $d = .28$. Lore in the firm is that promotion to partnership requires receiving nearly all 5s from supervisors in one’s yearly evaluations. We calculated the percentage of attorneys whose mean numerical evaluation was equal to or greater than 4.5. A higher percentage of men (14.00%) than women (4.76%) achieved evaluations in this top category, $\chi^2(1, N = 234) = 4.82$, $p < .03$.

Did this pattern of evaluations emerge in the narrative comments as well? No, judges’ ratings of the narratives on overall technical competence and interpersonal warmth did not differ by attorney sex ($M_{\text{techM}} = 3.50$, $SD = .48$, $M_{\text{techF}} = 3.45$, $SD = .43$, $M_{\text{warmM}} = 3.61$, $SD = .51$, $M_{\text{warmF}} = 3.59$, $SD = .41$), $t_s < 1$. However, evaluators used more positive performance words in their narratives about female attorneys ($M = 6.25$, $SD = 2.42$) than male attorneys ($M = 5.66$, $SD = 2.01$), $t(232) = -1.98$, $p < .05$. Thus, as predicted, numerical ratings revealed pro-male bias, whereas the narrative comments were more positive overall for women or showed no sex difference in specific mentions of technical competence and interpersonal warmth.

Narrative coding also revealed more indications that partnership was a possibility for male ($M = .06$, $SD = .17$) than female ($M = .02$, $SD = .14$) attorneys, $t(228) = 2.61$, $p < .01$. Looking at the data categorically, only 2 women (2.38%) and 2 men (1.33%) were explicitly labeled as “not partner material” by at least one evaluator, but the likelihood of partnership was mentioned by at least one evaluator for 5.95% of female attorneys and 14.67% of male attorneys. This categorical difference between partnership mentions and nonmentions and negative mentions by attorney sex was significant, $\chi^2(1, N = 234) = 4.01$, $p < .01$.

Correlations Among Judgments

We next computed correlations among the numerical ratings and narrative-coded attributes, separately for female and male attorneys. These appear in Table 1; entries above the diagonal refer to female attorneys and those below to male attorneys.

In the case of male attorneys, numerical ratings were correlated with all of the other coded variables, which themselves

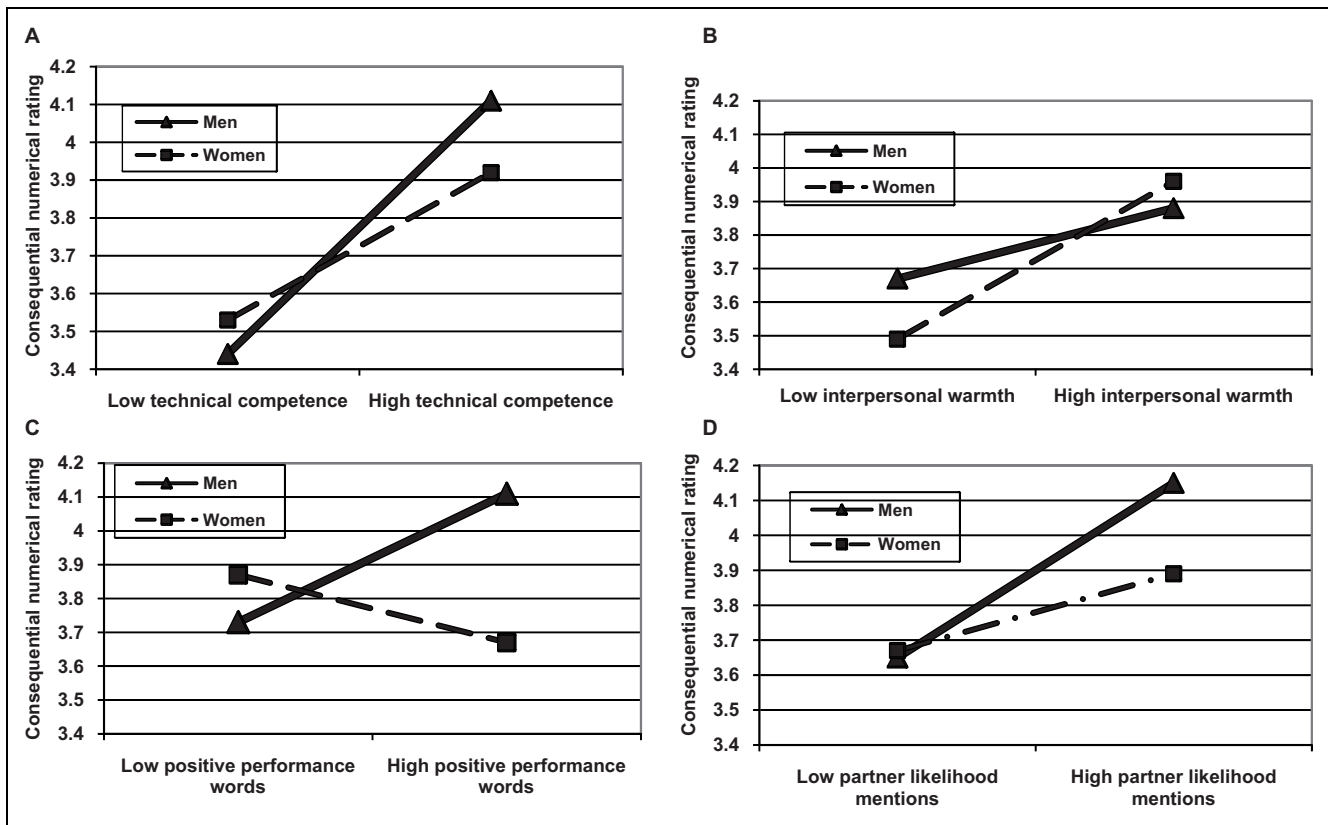


Figure 1. Interactions between attorney sex and narrative predictors on numerical ratings: Technical competence (A), interpersonal warmth (B), positive performance words (C), and partner likelihood (D)

were intercorrelated, suggesting a great deal of consistency in supervisors' judgments of men. For female attorneys, there was less consistency overall among these components, although narrative ratings of technical competence, interpersonal warmth, and partner likelihood did predict numerical ratings. Partner likelihood predicted positive narrative mentions of technical competence and interpersonal warmth for men but not for women. We examined apparent sex differences more explicitly by computing multiple regressions that included interactions between predictors and attorney sex.

Effects of narrative ratings of technical competence and interpersonal warmth on numerical ratings, by attorney sex. We regressed numerical ratings on attorney sex (centered), technical competence (centered), interpersonal warmth, and interactions between these predictors and attorney sex. Both interactions were significant: Attorney Sex \times Technical Competence, $B = .15$, $SE = .07$, $t(228) = 2.20$, $p < .05$, and Attorney Sex \times Interpersonal Warmth, $B = -.13$, $SE = .07$, $t(228) = -2.01$, $p < .05$.

We decomposed these interactions using utilities developed by Preacher, Curran, and Bauer (2006). These are depicted in Figure 1A (technical competence) and 1B (interpersonal warmth), using levels of the relevant predictor variable set at 1 *SD* above and below the mean. As can be seen in Figure 1A, narrative ratings of technical competence predicted numerical ratings for both female and male attorneys, $ps < .001$,

but the effect of technical competence was stronger for men than women. Additionally, the effect of attorney sex was not significant at low levels of perceived technical competence, $p > .30$ but was significant at high levels of perceived technical competence, $B = .10$, $SE = .03$, $t(228) = 2.83$, $p < .01$. That is, numerical ratings of male and female attorneys did not differ when attorneys were rated low on the narrative measure of technical competence, but numerical ratings were higher for men than women when narrative technical competence was high.

Figure 1B indicates that narrative ratings of interpersonal warmth predicted numerical ratings for both female and male attorneys, $ps < .001$, but warmth mattered significantly more for women. Furthermore, the effect of attorney sex was significant at low levels of warmth, $B = .09$, $SE = .01$, $t(228) = 6.96$, $p < .001$, but not at high levels of warmth, $p > .25$: Among attorneys described as low in interpersonal warmth, women received lower numerical ratings than men, but no sex difference emerged among attorneys described as high in interpersonal warmth.

Positive performance word effects on numerical ratings by attorney sex. The interaction between attorney sex and positive emotion words was also significant, $B = .04$, $SE = .02$, $t(229) = 2.10$, $p < .05$ and is plotted in Figure 1C. For men, positive performance words signaled positive numerical ratings, $B = .05$, $SE = .02$, $t(229) = 2.38$, $p < .02$, but for women they did not,

$p > .20$. Additionally, the sex difference in numerical ratings was significant when attorneys were described with high numbers of positive performance words, $B = .22$, $SE = .08$, $t(229) = 2.84$, $p < .01$, but not when they were described with few positive performance words, $p > .35$.

Partner likelihood effects on numerical ratings and attorney sex. Regression analysis also confirmed a significant Attorney Sex \times Partner Likelihood interaction on numerical ratings, $B = .45$, $SE = .21$, $t(228) = 2.17$, $p < .01$ (see Figure 1D). Partner likelihood predicted ratings for both sexes, $ps < .01$, but the effect was stronger for male attorneys. At low levels of partner likelihood, the effect of attorney sex was not significant, $p > .70$, but at high levels of partner likelihood, men received better numerical ratings than women, $B = .13$, $SE = .04$, $t(228) = 3.61$, $p < .001$. We also examined this pattern by comparing numerical ratings received by attorneys for whom at least one evaluator mentioned that partnership was likely. The 5 women who received any mention of partnership potential achieved numerical ratings of $M = 4.27$, $SD = .27$, whereas the mean numerical rating for the 22 partner-identified men was $M = 4.46$, $SD = .32$, $d = .43$.

Other variables. Other apparent differences in correlational patterns for female and male attorneys (Table 1) were not supported in multiple regressions testing interactions. For example, narrative technical competence and interpersonal warmth did not differentially predict partner likelihood mentions for women and men.

Discussion

Given the masculine context of finance law, we predicted that male attorneys would receive more favorable numerical ratings than female attorneys in this Wall Street firm. This was indeed the case, and based on the requirement of high ratings for partnership, more men than women in this firm were headed toward partnership. Of course, this finding alone tells us little about the presence of gender bias, for it is possible that men were objectively “better” performers than women. But arguing against this possibility is the fact that the pro-male bias was not apparent in narrative comments, which revealed greater reference to positive performance words for women than men, and no sex differences in narrative ratings of technical competence and interpersonal warmth. If men were objectively better, one might expect to see evidence of this in all forms of evaluation. Furthermore, when even the best men and women were compared—those judged equally highly favorably in narrative mentions of partner likelihood and technical competence—men still fared better in terms of the numerical ratings they received (more on this below).

Other findings in the performance evaluation literature have pointed to relatively low correlations between types of workplace evaluation and to different patterns of group differences depending on the type of evaluation (Roth, Huffcutt, and Bobko, 2003; Wilson, 2010). The precise cause of the different patterns of attorney sex effects cannot be specified with these data, but we suggest that gender stereotypes led to numerical

judgments that favored men and to the use of shifting standards in narrative language (Biernat, 2003). To the extent that men’s interpersonal warmth and women’s technical competence were considered relative to lower within-sex standards, stereotypical patterns of judgment may have been masked in open-ended language.

But it is also possible that heightened positivity toward women in narrative comments could reflect a kind of motivated “softening” of the message delivered in numerical ratings. Evaluators may have wanted to “throw a bone” to women, making their relatively negative numerical ratings appear less harsh (whether this harshness was driven by stereotyping or objective performance). They may also have been concerned about appearing prejudiced and sought to amplify their praise of women accordingly, describing them with more positive performance words and with assessments of technical skills and interpersonal warmth equaling that of men (see Harber, 1998, for a race-based example). But because junior attorneys did not have direct access to the narrative comments, we think it unlikely that motivated “softening” occurred, unless this pose was for the benefit of the other supervisors.

Instead, we suggest that gender stereotypes led to pro-male bias on the evaluative judgments that mattered, but that narrative references to technical competence and performance reflected the use of lower standards for women, which therefore muted or reversed the pro-male bias in numerical ratings.

Evidence of gender stereotyping is further suggested by the differential patterns of correlation between numerical ratings and other judgments for female and male attorneys. In general, there was more consistency between numerical ratings and attributes coded from the open-ended narratives for male than female attorneys. And though narrative interpersonal warmth and technical competence were correlated with numerical ratings for both female and male attorneys, multiple regression analysis indicated a sex-typed pattern: Technical competence mattered more for numerical ratings of men than women, and interpersonal warmth mattered more for numerical ratings of women than men.

Other research points to the importance of likeability in judgments of women in the workplace (Sidanius & Crane, 1989). For example, women who are described as “successful managers” nonetheless suffer evaluation decrements relative to men because they are perceived as lacking warmth (Heilman, Block, & Martell, 1995; Heilman, Wallen, Fuchs, & Tamkins, 2004). Our data also indicate that women’s numerical ratings suffered, relative to men’s, when they received low assessments of warmth. However, women did not particularly benefit relative to men when they were judged as high in warmth: Numerical ratings were roughly equal for “high-warmth” women and men. Thus, lacking warmth harmed women but having warmth did not offer them any greater benefit than it did men.

Also consistent with gender stereotyping patterns, narrative perceptions of technical competence mattered more for the numerical ratings of male than female attorneys. Partner likelihood was also higher for men than women, and mention of

partnership better predicted positive numerical ratings of male attorneys. What distinguished the “male competence” connection from the “female warmth” connection was that for men, receiving a low narrative assessment of technical competence did not harm them relative to women, but receiving narrative comments indicating high technical competence benefited them relative to women. That is, while both men and women may have been held to stereotyped expectations, women particularly suffered when they did not meet expectations, and men gained when they did.

These data also suggest that for women who were described in narratives as having high levels of perceived technical competence (and as partner material), numerical evaluations still did not match those of comparable men. For example, women described in the narratives as equivalently high in technical competence to men (e.g., at 1 *SD* above the mean) received numerical ratings of about half a *SD* lower than men’s. This difference could easily reduce partnership likelihood for women.

We also found that greater use of positive performance words in narrative descriptions of women—the only mean-level outcome that favored women—had no impact on the numerical ratings women received. However, positive performance word use for men did predict positive numerical ratings. This suggests that the positive narrative characterizations of women were more apparent than real; narrative comments about women may have reflected inflated positivity relative to low group level expectations—which did not translate into the consequential numerical ratings that matter for promotion.

This pattern of findings is reminiscent of Castilla’s (2008) performance–reward bias effect, whereby performance evaluations matter less for salary increases for women and minorities than for White men. Castilla argues that “employers consciously or unconsciously discount the performance ratings of employees because of their gender, race, or nationality,” when it comes to making salary decisions (p. 1483). We are doubtful that explicit discounting played a role in our data, as narrative comments were generated at the same time as numerical ratings, by the same evaluators. In Castilla’s data, evaluations were generated in one stage and wage determinations made in a second stage, often by different evaluators. Still, stereotype-based expectations underlie both effects. Of course, we did not examine the effects of evaluations on salary increases. But in this law firm, close-ended ratings are the outcomes that matter; only attorneys who receive mainly 5s are headed for partnership. Men are clearly favored in these judgments, and these judgments are more directly tied to the positive narrative comments about performance that men than women received.

Overall, our data indicate that male attorneys received evaluative advantages at this firm. While the absolute sex difference in numerical ratings (*M*s of 3.90 vs 3.76 on 5-point scale) may not seem very large, the firm’s reliance on this number for partnership consideration makes it nearly three times more likely that men than women will be promoted to partner. Again, it is possible that this difference reflects objective performance differences rather than gender bias, but we think this is unlikely for several reasons. First, narrative

content did not support the numerical rating difference—if men were objectively “better,” this should have been evident in all forms of evaluation. Second, though men and women were perceived equivalently in technical competence and interpersonal warmth, these factors predicted consequential numerical ratings in sex-typed ways. Such patterns cannot be explained by objective performance differences. Third, even when one considers the most highly regarded women—those who were “matches” to men in terms of narrative description of high technical competence and partnership material—their numerical ratings did not match those of men.

These data provide a snapshot of the evaluation process of one law firm at one point in time, and our analyses suggest that gender affected the evaluations junior attorneys received. The real-world context did not offer us the tight control of the lab, in which employee sex could be isolated as a causal factor in evaluative bias and process could be assessed. But this setting provided a valuable venue for testing predictions about subtle and complex forms of gender-stereotyping and for considering consequences for real employees being judged by real evaluators in real time.

Acknowledgments

We are grateful to Hillary Hansen, Jennifer Takehana, Jay Middleton, and Megan Geimer for their assistance in coding open-ended narratives.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this paper was facilitated by NSF grant #51720 awarded to the first author.

Notes

1. The number of male supervisors who offered evaluations predicted the overall numerical ratings attorneys received, $r = .19, p < .01$. This effect was not moderated by attorney sex and did not moderate any of the substantive findings described in the text; therefore, it will not be discussed further.
2. Although coders judged partner likelihood categorically, we treated this as a continuous measure because we averaged across all such mentions received by an attorney. For example, if an attorney was evaluated by six supervisors, and one mentioned the attorney was headed for partnership but the others said nothing about partnership, the resulting “partner likelihood” score was .167.

References

- Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58*, 1019-1027.
- Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin, 36*, 855-868.

- Biernat, M., & Vescio, T. K. (2002). She swings, she hits, she's great, she's benched: Implications of gender-based shifting standards for judgment and behavior. *Personality and Social Psychology Bulletin*, 28, 66-77.
- Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology*, 48, 587-605.
- Castilla, E. J. (2008). Gender, race, and meritocracy in organizational careers. *American Journal of Sociology*, 113, 1479-1526.
- Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, 92, 631-648.
- Dipboye, R. L. (1985). Some neglected variables in research on discrimination in appraisals. *The Academy of Management Review*, 10, 116-127.
- Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109, 573-598.
- Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82, 878-902.
- Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality & Social Psychology*, 70, 491-512.
- Harber, K. (1998). Feedback to minorities: Evidence of a positive bias. *Journal of Personality & Social Psychology*, 74, 622-628.
- 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., Block, C. J., & Martell, R. F. (1995). Sex stereotypes: Do they influence perceptions of managers? *Journal of Social Behavior and Personality*, 10, 237-252.
- Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: Reactions to women who succeed at male-stereotyped tasks. *Journal of Applied Psychology*, 89, 416-427.
- Heneman, R. L. (1986). The Relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology*, 39, 811-826.
- Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, 89, 899-913.
- Kite, M. E., Deaux, K., & Haines, E. L. (2007). Gender stereotypes. In F. L. Denmark & M. A. Paludi (Eds.), *Psychology of women: Handbook of issues and theories* (2nd ed. pp. 205-236). Westport, CT: Praeger Publishers/Greenwood Publishing Group.
- Kraiger, K., & Ford, J. K. (1990). The relation of job knowledge, job performance, and supervisory ratings as a function of race. *Human Performance*, 3, 269-279.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.
- 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.
- Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.
- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31, 437-448.
- Roth, P. L., Huffcutt, A. I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology*, 88, 694-706.
- Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, 74, 629-645.
- Rudman, L. A., & Fairchild, K. (2004). Reactions to counterstereotypic behavior: The role of backlash in cultural stereotype maintenance. *Journal of Personality and Social Psychology*, 87, 157-176.
- Rudman, L. A., & Glick, P. (2008). *The social psychology of gender: How power and intimacy shape gender relations*. New York, NY: Guilford Press.
- Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology*, 19, 174-197.
- Spence, J. T., & Buckner, C. E. (2000). Instrumental and expressive traits, trait stereotypes, and sexist attitudes. *Psychology of Women Quarterly*, 24, 44-62.
- Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, 16, 474-480.
- Vescio, T. K., Gervais, S. J., Snyder, M., & Hoover, A. (2005). Power and the creation of patronizing environments: The stereotype-based behaviors of the powerful and their effects on female performance in masculine domains. *Journal of Personality and Social Psychology*, 88, 658-672.
- Wilson, K. Y. (2010). An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations*, 63, 1903-1933.

Bios

Monica Biernat is a Professor of Psychology at the University of Kansas.

M. J. Tocci is a former Deputy District Attorney for Alameda County, California, and currently works as a consultant and teacher of trial advocacy skills.

Joan C. Williams is a Distinguished Professor of Law at the University of California Hastings College of the Law, and director of the Center for WorkLife Law.