Sex Roles, Vol. 47, Nos. 9/10, November 2002 (© 2002)

Reducing the Effects of Gender Stereotypes on Performance Evaluations[1]

Cara C. Bauer[2,3] and Boris B. Baltes[2]

The purpose of this research was to extend previous work on gender bias in performance evaluation. Specifically, we examined whether a structured free recall intervention could decrease the influence of traditional gender-stereotypes on the performance evaluations of women. Two hundred and forty-seven college students provided performance ratings for vignettes that described the performance of male or female college professors. Results indicated that without the intervention, raters who have traditional stereotypes evaluated women less accurately and more negatively. Conversely, the structured free recall intervention successfully eliminated these effects. The usefulness of the structured free recall intervention as a tool for decreasing the influence of gender stereotypes on performance ratings is discussed.

KEY WORDS: performance evaluation; stereotypes; gender bias; structured free recall.

A large amount of the research on performance appraisal concerns the effect of ratee gender on performance evaluations and other merit ratings (e.g., Arvey, 1979; Davison & Burke, 2000; Deaux & Taynor, 1973; Dobbins, Cardy, & Truxillo, 1988; Goldberg, 1968; Gunderson, Tinsley, & Terpstra, 1996; Hamner, Kim, Baird, & Bigoness, 1974; Martell, 1996; Maurer & Taylor, 1994; Mobley, 1982; Pazy, 1986; Pulakos, White, Oppler, & Borman 1989; Robbins & DeNisi, 1993; Shaw, 1972; Sidanius & Crane, 1989; Swim, Borgida, Maruyama, & Myers, 1989; Thompson & Thompson, 1985; Yammarino & Dubinsky, 1988). The majority of the research has focused on whether or not a pro-male bias exists and what possible causes of the bias might be. According to Nieva and Gutek (1980), a pro-male bias occurs when men are rated more favorably than women given similar performance. Although the research, especially in field settings, has yielded mixed results, there does seem to be some evidence for a pro-male bias in the evaluation of performance and in employment hiring decisions (Arvey, 1979; Davison & Burke, 2000; Deaux & Taynor, 1973; Dobbins et al., 1988; Goldberg, 1968; Gunderson et al., 1996; Martell, 1996; Maurer & Taylor, 1994; Pazy, 1986; Robbins & DeNisi, 1993; Shaw, 1972; Sidanius & Crane, 1989). If this is the case, the implications for women in organizations are troubling as women may still be at a disadvantage in obtaining employment, pay increases, and promotions.

The majority of past performance appraisal research has focused specifically on the gender of the ratee as the cause of the differences in ratings, in other words a main effect of ratee gender (e.g., Deaux & Taynor, 1973; Goldberg, 1968; Gunderson et al., 1996; Pazy, 1986; Sidanius & Crane, 1989), however some researchers have looked for interactions with social cognitive variables (e.g., stereotypes, prejudices) for answers to this question (e.g., Dobbins et al., 1988; Martell, 1996; Maurer & Taylor, 1994; Robbins & DeNisi, 1993). These researchers (Dobbins et al., 1988; Maurer & Taylor, 1994; Robbins & DeNisi, 1993) have applied gender theory to explain the cause of the pro-male bias found in some performance evaluations. It is thought, and research has shown (e.g., Dobbins et al., 1988), that only raters with traditional stereotypes of women will exhibit a pro-male bias.

[1] Portions of this paper were originally presented at the 16th annual meeting of the Society for Industrial and Organizational Psychology, San Diego, California, April 2001.

[2] Wayne State University, Detroit, Michigan.

[3] To whom correspondence should be addressed at Psychology Department, 71 W. Warren, Wayne State University, Detroit, Michigan 48202; e-mail: cbauer@sun.science.wayne.edu.

0360-0025/02/1100-0465/0 © 2002 Plenum Publishing Corporation

Furthermore, recent research that has examined the cognitive components of performance appraisal may provide the means by which one could reduce the pro-male bias exhibited by raters with traditional stereotypes. For example, Martell (1996) examined the functioning of heuristics based on gender stereotypes and their effects on evaluations of women. Martell found that gender stereotypes may cause raters to ascribe more effective work behaviors to men than women; a systematic response bias (rather than selective memory) caused the differences. This systematic response bias is thought to occur because the rater relies on a heuristic/stereotype at the time of the rating instead of thinking back to the actual performance (Baltes & Parker, 2000a, 2000b; Martell, 1996). Recently, a structured free recall intervention has been shown to reduce raters' reliance on heuristics and increase rating accuracy (Baltes & Parker, 2000b). The purpose of the present study was to extend previous research on gender bias in performance evaluation by testing a structured free recall intervention to reduce the influence of a pro-male bias in performance ratings.

Pro-Male Bias

The inconsistent results in the literature on performance appraisal have resulted in a controversy about the existence of a pro-male bias. Although many researchers have found support for the notion that men are, in general, rated more favorably than women (e.g., Deaux & Taynor, 1973; Goldberg, 1968; Gunderson et al., 1996; Pazy, 1986; Shaw 1972; Sidanius & Crane, 1989), some researchers have failed to find a significant difference between the ratings provided for men and women (e.g., Hamner et al., 1974; Mobley, 1982; Pulakos et al., 1989; Swim et al., 1989; Thompson & Thompson, 1985; Yammarino & Dubinsky, 1988).

The inconsistency of findings of the pro-male bias in performance evaluations is problematic. However, some researchers have offered several explanations as to why these inconsistencies may exist. For example, research has shown that women will only be discriminated against when the job type is viewed as being traditionally a men's job (see Martinko & Garner, 1983; Kalin & Hodgins, 1984, for reviews). Also, researchers have shown that the type and amount of information given to a rater can affect the strength of gender stereotypes on subsequent ratings (Davison & Burke, 2000, Fiske, 1991; Glick, Zion, & Nelson, 1988; Locksley, Borgida, Brekke, & Hepburn, 1980; Locksley, Hepburn, & Ortiz, 1982; Pratto & Bargh, 1991, Seta & Seta, 1993). The more information presented, and the more the information refutes the stereotype, the less raters apply the stereotype to the particular person in question. Another hypothesis more pertinent to the present study is the idea that most researchers have assumed that all or at least most raters will exhibit biased behavior, which may be false. Social psychologists have long believed that pro-male bias lies in the schema or stereotype people have about women. If this were the case, then whether support for the pro-male bias was found would depend on whether or not the raters in the study believed in the stereotype. Because the previously mentioned performance evaluation studies were only assessing ratee gender and not the raters' stereotypes, this factor could also account for some of the mixed results.

Gender-Based Stereotypes

According to social–cognitive theory, most raters have well-developed stereotypes of men and women (Bem, 1981; Del Boca, Ashmore, & McManus, 1986; Fishbein & Ajzen, 1975; Swim & Sanna, 1996), which link men and women to certain behaviors and characteristics (Del Boca & Ashmore, 1980). For instance, according to Del Boca and Ashmore (1980) the positively valued characteristics of the male stereotype include competence, rationality, and assertion; whereas the positively valued female characteristics include warmth and expressiveness. In a nutshell, "the typical woman is seen as nice but incompetent, the typical man as competent but maybe not so nice" (Fiske, 1998, p. 377).

Stereotyping involves generalizing beliefs about groups as a whole to members of those groups. When they rely on stereotypes people can categorize others into groups on numerous demographic bases, including gender, religion, and race, and perceptions of specific individuals will be influenced by what people think they know about the group as a whole (Cleveland, Stockdale, & Murphy, 2000). Cleveland and her colleagues stated that gender stereotypes are socially shared beliefs about the characteristics or attributes of men and women in general that influence our perceptions of individual men and women.

Bem (1981) stated that there are individual differences in people's gender role stereotypes, in that some individuals are more gender-typed and hold to more traditional beliefs that women are dependent,

illogical, and ineffective. Gender-typed individuals view maleness and femaleness as two separate categories, and they rely on their schemata about these notions to evaluate and organize information. On the other hand, individuals who are nongender-typed do not rely to the same extent on gender stereotypes to organize information.

These notions have relevance for performance evaluation. That is, the differences between ratings of men and women may be a consequence of the raters' social–cognitive processes rather than the sex of the rater. Further, because ratings are biased in the direction of the characteristics of the stereotype, the individual differences in raters' stereotypes of women may be biased (DeNisi, Cafferty, & Meglino, 1984). Therefore, raters who hold a traditional stereotype of women will associate women with ineffectiveness and will too often attribute ineffective performance to women. Conversely, raters who believe less strongly in the traditional stereotype of women will not associate women with ineffective performance, and will not overly attribute low performance behaviors to women. In other words, those who do not endorse traditional stereotypes do not view the performance of women as being less effective than that of men and thus do not use a person's gender to aid in the organization/classification of performance information.

Several studies have been conducted to examine the notion that gender-based stereotypes bias performance evaluations (Dobbins et al., 1988; Martell, 1996; Maurer & Taylor, 1994; Robbins & DeNisi, 1993), and all have shown support for this theory. Dobbins et al. (1988) examined the effects of individual differences on the evaluation of hypothetical male and female professors of varying levels of performance. The undergraduates who participated in this study completed the WAPS (Women as Professors Scale) in order to assess their stereotypes of women in the specific role of a professor. If the participant received a low score on the WAPS it was thought to indicate a traditional stereotype of women as professors (i.e., associated women with ineffectiveness). High scores indicated that the individuals do not believe in (or had less strong beliefs in) traditional stereotypes (i.e., women professors were not associated with low performance). In addition, the participants rated the performance of either four fictitious male professors or four fictitious female professors. Dobbins et al. (1988) manipulated ratee gender as a between subjects factor in order to avoid alerting participants that differences between the ratings of men and women were being examined. The researchers found that ratee gender significantly moderated the relationship between stereotype (WAPS scores) and accuracy (as measured by both differential elevation accuracy and average deviation accuracy) of evaluations. In other words, those with a traditional stereotype rated women professors lower than men and rated them less accurately.

Dobbins et al. (1988) argued that underlying gender differences in performance evaluations are a result of the social–cognitive processes of raters. Individuals who hold traditional stereotypes of women often perceive women's performance as an outcome of unstable or situationally derived factors and, as a result, judge it as less worthy and less stable. The raters who used a biased information-processing strategy might have been more likely to associate low performance behaviors with women, which result in inaccurate ratings. This notion is further supported by Martell (1996), who stated that stereotypes lead to differential performance expectations (positive for men and negative for women), which in turn lead to more favorable performance ratings for men than women. Therefore, those with a traditional stereotype of women may rely on the stereotype as a heuristic at the time of the rating, which may result in lower performance ratings for women than men. Maurer and Taylor (1994) replicated and extended the work of Dobbins et al. (1988) with similar results.

Other researchers have found support for the notion that people who have expectations about another individual's performance may adopt a biased decision criterion when rating the person's behaviors (Baltes & Parker, 2000b; Martell; 1996; Martell & Guzzo, 1991; Martell & Willis, 1993). According to Baltes and Parker (2000b), these raters are not relying on their memory, but rather on how prone they are to endorse behaviors that are consistent with their performance expectations and to reject behaviors that are not consistent with their expectations. In other words, when a performance expectation exists, raters evaluate the behaviors that are consistent with their corresponding beliefs as more likely to occur.

Reducing Gender-Based Stereotypes

If some raters have traditional gender stereotypes, which result in a biased decision criterion, we must find a method to correct the problem. It is not feasible to propose that all managers with traditional stereotypes be stopped from making evaluations of women. Therefore, an intervention is necessary in

order to reduce the negative effects of gender stereotypes on performance ratings. Social psychology researchers have examined the issue of stereotypes and stereotype reduction quite extensively. The simplest way researchers have found to control stereotypes is to provide information about the target that is inconsistent with stereotypes (for a review, see Fiske & Neuberg, 1990). However, this type of "intervention" cannot realistically be used in real world performance rating situations. That is, we cannot provide stereotype inconsistent information to a recruiter about to interview a potential employee. Other interventions, which may be more applicable to real world performance rating situations, include trying to force raters to control their stereotype or motivate raters to be more accurate. With respect to the first idea, researchers have attempted to force participants to control their stereotypes (Bodenhausen & Macrae, 1996; Wegner, 1994). However, these studies have not often been successful and have actually led to rebound effects where the stereotypes become more powerful than before (Macrae, Milne, & Bodenhausen, 1994). Researchers have also attempted to motivate participants to be more accurate (and thereby to discriminate less) by striving for accuracy and/or by rewarding accuracy (Nelson, Acker, & Manis, 1996; Neslon, Biernat, & Manis, 1990; Snyder, Campbell, & Preston, 1982). These attempts have had mixed results, which led Fiske (1998) to suggest that motivation, although helpful in reducing the impact of stereotypes, cannot cure stereotyping, prejudice, or discrimination. In summary, it seems that most previous research interventions are either not applicable to real world performance rating situations or have not worked consistently in reducing stereotypes. However, another intervention (structured free recall) based on social cognition research that has been recently found to be successful in reducing the influence of externally generated stereotypes (Baltes & Parker, 2000b) may also be successful in reducing gender biases.

In free recall interventions, raters are instructed to recall behaviors that they have observed and to rely on those observations to complete the rating. This should reduce the raters' reliance on judgments, which are influenced by stereotypes, to make their performance ratings. Feldman and Lynch (1988) stated that the accessibility and diagnosticity of information in memory can influence evaluative responses. Diagnosticity refers to whether or not previous judgments or stored information are perceived to be relevant to subsequent judgments, whereas accessibility refers to the ease with which a cognitive construct can be brought into awareness. In other words, accessibility and diagnosticity influence whether a prior cognition will be used as an input to a related judgment. According to Baltes and Parker (2000a), by recalling behaviors that were displayed by the ratee before the performance evaluation, raters should increase the accessibility of these specific memories and increase the likelihood of using these memories when they rate the performance. In essence, a free recall intervention is an attempt to reduce a rater's reliance on an overall judgment of the ratee (which is often biased by stereotypes) by getting the rater to use specific observed behaviors to complete the performance ratings.

Baltes and Parker (2000b) examined the role of a structured free recall intervention in reducing the effects of performance expectations on behavioral ratings. They focused specifically on reducing the "performance cue effect" through a free recall intervention. Performance cues can consist of any overall performance information that has been obtained from sources such as prior employers, employee resumes, and/or prior interviewers. Further, people's ratings can be affected by such performance feedback. Thus, the performance cue effect (PCE) is a phenomenon where expectations of performance can cause a cue-consistent bias in ratings. This cue-consistent bias is caused by heuristic use on the part of the raters. It is similar to the rating bias caused by gender stereotypes. Baltes and Parker found that the structured free recall intervention successfully reduced the performance cue effect. Structured free recall is thought to improve ratings by reducing raters' reliance on heuristics and increasing their reliance on observed behaviors. Therefore, a structured free recall intervention strategy may also be able to remove internal biases like gender stereotype bias from performance ratings.

It should be pointed out that the theory behind the structured free recall strategy is supported by prior social psychology research that has shown that individuating information can reduce reliance on stereotypes (Glick et al., 1988; Locksley et al., 1980, 1982; Pratto & Bargh, 1991; Seta & Seta, 1993). That is, the information recalled by raters could act as information that will reduce raters' reliance on their stereotypes. The difference between the structured free recall intervention and previous research is that the information will be internally generated by the rater instead of being presented to the rater by the experimenter. This is especially important because the only way for an intervention to work in an applied setting is for it not to be dependent on information

provided by the experimenter, which, as stated above, is typical of previous research on reducing the influence of stereotypes.

The Present Study

Following Dobbins et al. (1988), the present study was designed to examine the effects of individual differences in stereotypes of women on performance ratings of college professors. The role of a professor was chosen as the stimulus profession for several reasons. First, we attempted to replicate and extend the research of Dobbins et al. (1988) and, in doing so, attempted to maintain a similar experimental design. Second, we collected data from undergraduates, and the role of a college professor is one that they can rate with some degree of knowledge and experience.

Lastly, we investigated whether a structured free recall intervention strategy can remove the impact of gender stereotypes from performance ratings. If structured free recall can reduce the impact of other cognitive processes (e.g., performance cues), it is possible that it may also effectively reduce the impact of gender stereotypes on the accuracy and mean level of performance evaluations. Specifically, a structured free recall intervention should successfully reduce the bias in ratings exhibited by raters with traditional stereotypes of women. Therefore, the hypotheses of the present study are as follows: (1) The impact of gender stereotypes on the mean level of performance evaluations will be diminished by a structured free recall intervention; (2) the impact of gender stereotypes on the accuracy of performance evaluations will be diminished by a structured free recall intervention.

METHOD

Participants

Two hundred sixty-five undergraduate students (180 women and 85 men) at a large Midwestern university were recruited to participate in the study. They participated for extra credit in Introductory Psychology courses. Although specific information about age and racial/ethnic group was not collected, the make up of the psychology 1010 classes at the university, from which our sample came, is as follows: mean age = 20.10 years; racial/ethnic background is Black/African American = 28.2%, White = 50.0%, Hispanic = 2.6%, Arabic/Middle Eastern = 5.4%, Asian = 6.0%, and Multiracial = 2.8%.

Materials

The Women as Professors Scale (WAPS)

The WAPS, a 21-item scale, was developed by Dobbins et al. (1988) and is a revision of the Women as Managers Scale (WAMS; Peters, Terborg, & Taynor, 1974). The WAP scale specifically assesses people's stereotype of women in the specific role of a college professor. The coefficient alpha for the WAPS in the present study was .89.

Vignettes

Two vignettes were constructed for use in this study. The two vignettes are similar to vignettes constructed by Dobbins et al. (1988) and Cardy and Kehoe (1984). Each vignette consisted of 20 incidents of classroom behavior that were chosen from Sauser, Evans, and Champion's (1979) 250 scaled incidents of college classroom teaching behavior. The 20 behaviors correspond to five dimensions of instructor performance (ability to present material, interest in course and material, relationship with students, reasonableness of workload, and fairness of testing and grading); four behaviors from each dimension were selected to form the vignette. The vignettes were constructed so that ratees would exhibit a range of performance levels across dimensions. That is, each ratee performed well on some dimensions and average or poorly on others. In order to determine a true score (e.g., the expected ratee performance level) for each vignette we used the same method as Dobbins et al. (1988) and obtained overall true scores of 7.18 and 6.75 for the two vignettes. The 20 critical incident behaviors used to construct the vignettes were pretested in a group of 18 graduate student subject-matter experts and virtually identical true scores were obtained. For an example of one of the vignettes, see the Appendix.

Performance Rating Scales

The rating measure consisted of 10 11-point behaviorally anchored items. The scales were constructed to measure the five dimensions of instructor

performance, and each dimension was measured with two individual items. They were based on the scaled critical incidents developed by Sauser et al. (1979) and were similar to those used by Dobbins et al. (1988). The reliabilities for the two item dimensional scales ranged from .58 to .82.

Procedure

The study was conducted in two phases. First, all participants completed an instrument designed to measure their stereotypes of female professors (the WAP scale) during a mass testing of all students enrolled in Introductory Psychology. Students were then invited to participate in the second phase of the study based on their scores on the mass testing stereotype measure. In the second phase of the study, participants were assigned to one of four conditions: female ratees/no structured free recall; male ratees/no structured free recall; female ratees/structured free recall; and male ratees/structured free recall. After signing the consent form, the participants were given two vignettes of teaching performance that described the behavior of either male or female professors at different levels of performance. The behavior of the men and women in the vignettes was identical, except that gender-relevant terminology was switched.

The information given to the participants about the nature of the ratings was identical to that given by Dobbins et al. (1988). The participants were informed that they were participating in a project designed to identify outstanding teachers and that the ratings would be used by the university "to make personnel decisions affecting the professor's salary, promotion, and tenure status within the university" (p. 553).

After listening to the instructions from the experimenter, the participants studied the first vignette for 5 min. Next, the participants in the structured free recall condition were given 5 min to recall as many positive behaviors as possible that were relevant to the performance dimensions on which they rated the professor. They were then given 5 min more to recall as many negative behaviors as possible (the order of positive and negative behaviors was counter balanced). The participants were asked to record the behaviors on a sheet of paper that listed the five dimensions of instructor performance at the top, and they were then encouraged to refer back to their recall sheets when they completed the performance evaluation. The participants in the control condition followed the same sequence except that, instead of the structure free recall intervention, they completed half of the following distracter tasks before the first vignette and the remainder before rating the second vignette in order to account for the difference in time. The distracter tasks included: Snyder and Gangestad's (1986) 18-item Self-Monitoring Scale; Cacioppo and Petty's (1982) Need for Cognition Scale; Watson, Clark, and Tellegen's (1988) PANAS Scale; Macdonald's (1970) AT-20 Tolerance for Ambiguity Scale; and Webster and Kruglanski's (1994) Need for Closure Scale.

Next, participants in all conditions were given a rating scale with which they rated the performance of the professors. This procedure was then repeated for the second vignette. After completing the second vignette, they were given a manipulation check to assess the gender of the ratee manipulation. Finally, upon completion of all measures, the participants received a verbal debriefing discussing the purpose of the study.

RESULTS

Manipulation Check

Participants were asked to indicate whether or not each of the two professors was male or female. Ninety-three percent of the participants correctly reported the gender of the instructors in the vignettes. Though a 100% impact was not obtained, the gender of the ratee appears to have been manipulated successfully. These results are similar to those of Dobbins et al. (1988), where 88% correctly reported the gender of the instructor.

Mean Level and Accuracy of Ratings

To test the hypotheses, hierarchical regression analyses were performed separately on the accuracy measures to examine the ratee gender by WAPS by intervention type interaction. Each accuracy measure was regressed on gender of the ratee (man or woman), stereotype of women (WAPS score), the intervention type (structured free recall or no structured free recall), and all possible interactions. The nonstandardized and standardized regression weights obtained in these analyses are displayed in Tables I and II, along with the total amount of variance accounted for by the predictor variables. Participant gender was also originally included in all our analyses. However, no

Table I. Summary of Regression Analyses Predicting Average Deviation Scores

Variable	N	R^2	B	$SE\ B$	β
Model with all participants	247	.054			
WAPS × RG × I			−.407	.310	−.083
Control group	115	.071*			
WAPS Score (WAPS)			.134	.109	.114
Ratee Gender (RG)			−.200	.174	−.106
WAPS × RG			.515	.218	.217*
Structured free recall condition	132	.014			
WAPS Score (WAPS)			.009	.109	.073
Ratee Gender (RG)			−.174	.168	−.091
WAPS × RG			.107	.219	.043

Note. WAPS = Women as Professors Scale; RG = Ratee Gender; I = Intervention Type.
$^{*}p < .05$. $^{**}p < .01$ (two-tailed).

main or interaction effect of participant gender was found and thus it was dropped from the analyses.

Mean Level of Ratings

To assess the direction of error (either positive or negative) in the performance ratings, the average (unsquared) deviation between ratings and true levels of performance was computed. This measure was calculated by averaging the deviations across dimensions and ratees. Specifically, negative (positive) deviations indicate ratings that were lower (higher) than the true performance level. Hypothesis 1 states that the impact of gender stereotypes on the mean level of performance evaluations will be diminished by a structured free recall intervention. The results of the hierarchical regression analysis revealed that none of the main effects or two-way interactions were significant predictors of average deviation accuracy. The three-way interaction approached, but did not reach, significance ($p < .10$). The data were then split into the intervention types (control and structured free recall), and the average deviation scores were regressed on gender of the ratee (men or women), stereotypes of women (WAPS score), and the interaction between the two for each group separately. The nonstandardized and standardized regression coefficients obtained in this analysis are displayed in Table I, along with the total amount of variance accounted for by the predictor variables.

In the control condition, analyses indicated that the Ratee Gender × WAPS interaction significantly affected the average deviations of ratings, $t(111) = 2.36$, $p < .05$. As can be seen in Fig. 1, these results indicate that women were evaluated less favorably and with a larger negative bias by raters with more traditional stereotypes of women. However, participants' stereotype of women (i.e., WAPS score) was not significantly related to more or less favorable ratings when men were evaluated.

On the other hand, in the structured free recall condition, the Ratee Gender × WAPS interaction did not affect the average deviations of ratings (see Fig. 2). Results indicate that raters with traditional stereotypes of women did not evaluate women less favorably than did raters who believed less strongly in the traditional stereotype.

Table II. Summary of Regression Analyses Predicting Differential Elevation Scores

Variable	N	R^2	B	$SE\ B$	β
Model with all participants	247	.021			
WAPS × RG × I			1.186	.510	.148*
Control group	115	.121**			
WAPS Score (WAPS)			−.212	.119	−.159
Ratee Gender (RG)			.228	.191	.107
WAPS × RG			−.786	.239	−.294**
Structured free recall condition	132	.017			
WAPS Score (WAPS)			−.113	.215	−.046
Ratee Gender (RG)			−.325	.331	−.086
WAPS × RG			.401	.432	.081

Note. WAPS = Women as Professors Scale; RG = Ratee Gender; I = Intervention Type.
$^{*}p < .05$. $^{**}p < .01$ (two-tailed).

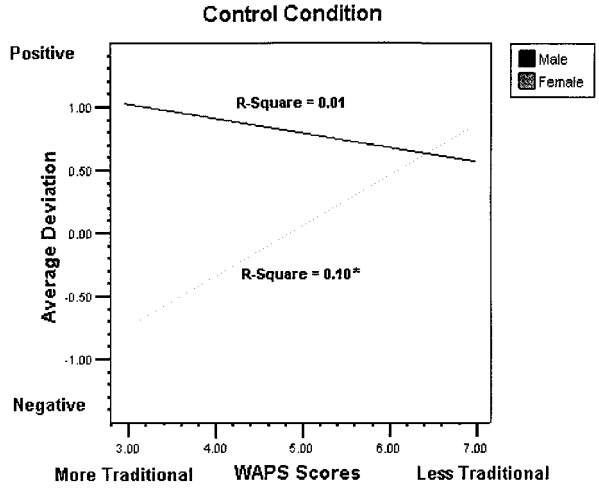

Fig. 1. Depiction of the two-way interaction between Ratee Gender and WAPS score on the average deviation accuracy measure in the control condition. *Note.* WAPS = Women as Professors Scale ($^{*}p < .05$, $^{**}p < .01$, two-tailed).

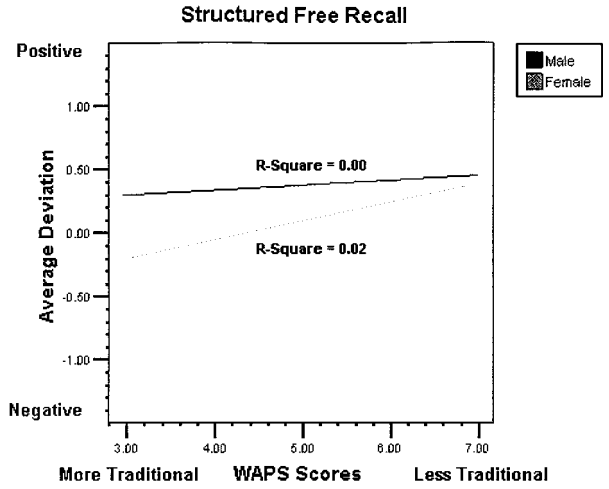

Fig. 2. Depiction of the two-way interaction between Ratee Gender and WAPS score on the average deviation accuracy measure in the structured free recall condition. *Note.* WAPS = Women as Professors Scale ($^{*}p < .05$, $^{**}p < .01$, two-tailed).

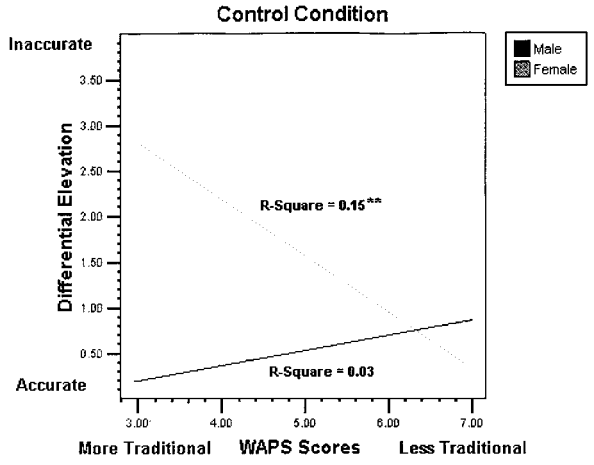

Fig. 3. Depiction of the two-way interaction between Ratee Gender and WAPS score on the differential elevation accuracy measure in the control condition. *Note.* WAPS = Women as Professors Scale ($^{*}p < .05$, $^{**}p < .01$, two-tailed).

Accuracy of Ratings

In order to replicate Dobbins et al. (1988), the accuracy of the participants' ratings was also measured using differential elevation, one of Cronbach's (1955) components of accuracy. Differential elevation (DEL) reflects the degree to which a rater differentiates between average ratee performance levels. This accuracy measure was calculated with formulae presented by Murphy, Garcia, Kerkar, Martin, and Balzer (1982). Large accuracy scores reflect inaccurate ratings, whereas small scores reflect accurate ratings. Hypothesis 2 states that the impact of gender stereotypes on the accuracy of performance evaluations will be diminished by a structured free recall intervention. The data support this prediction, as the three-way interaction between ratee gender, WAPS, and intervention type was significant, $t(245) = 2.33$, $p < .05$. None of the main effects or two-way interactions were found to be significant. To examine the three-way interaction, the data were then split into the intervention types (control and structured free recall), and the differential elevation scores were regressed on gender of the ratee (men or women), stereotypes of women (WAPS score), and the interaction between the two for each group separately. The nonstandardized and standardized regression coefficients obtained in this analysis are displayed in Table II, along with the total amount of variance accounted for by the predictor variables.

In the control group, the Ratee Gender × WAPS interaction was significant, thus this interaction significantly affected differential elevation, $t(111) = -3.29$, $p < .01$. As can be seen in Fig. 3, these results are consistent with those obtained by Dobbins et al. (1988) and indicate that raters with traditional stereotypes of women less accurately differentiated among average levels of women's performance. Specifically, women were evaluated less accurately by raters with low scores on the WAPS than by raters with high WAPS scores, however, scores on the WAPS were not significantly related to the accuracy with which men were evaluated. However, the Ratee Gender × WAPS interaction was not significant in the structured free recall condition (see Fig. 4). Raters with

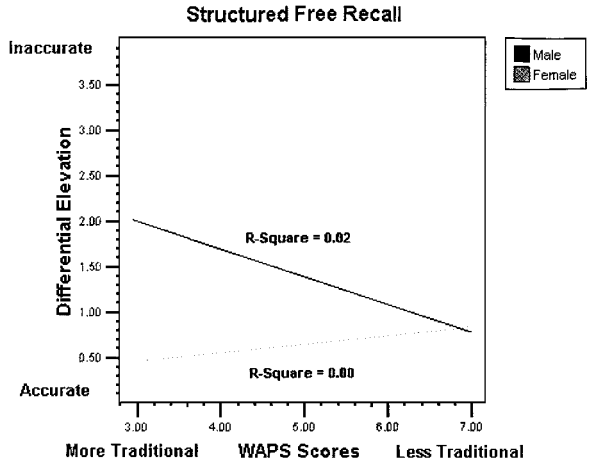

Fig. 4. Depiction of the two-way interaction between Ratee Gender and WAPS score on the differential elevation accuracy measure in the structured free recall condition. *Note.* WAPS = Women as Professors Scale ($^{*}p < .05$, $^{**}p < .01$, two-tailed).

strong traditional stereotypes of women differentiated among average levels of women's performance as accurately as other raters. Thus, it appears that the effects of gender stereotypes can be removed from performance accuracy measures through the implementation of structured free recall.

Overall, the data seem to support both of our hypotheses, as neither of the two-way interactions between Ratee Gender × WAPS were significant in the structured free recall condition. In other words, gender of the ratee no longer moderated the relationship between accuracy and stereotype in the structured free recall condition.

DISCUSSION

The present study had two main goals: (1) to illustrate the influence of bias on the evaluation of women; and (2) to test an intervention for reducing the impact of this bias. We accomplished these goals by measuring the raters' stereotype of women, and by testing a structured free recall intervention to reduce the impact of this stereotype. Further, the data provide support for both hypotheses. Female ratees were evaluated less accurately by raters with strong traditional stereotypes of women, in the control condition. This result is consistent with Dobbins et al. (1988), who stated that for female ratees who are evaluated by raters with traditional stereotypes of women, the distribution of rewards and sanctions based on a highest–lowest average performance-level criterion will be unequal to that of men, because the differential elevation component of rating accuracy was affected. The average deviation measure of accuracy was also affected, which indicates that individuals with traditional stereotypes evaluate women lower than their true performance level. This suggests that when appraisals are used to make administrative decisions or distribute merit pay, women who are evaluated by raters with traditional stereotypes may receive less positive outcomes than their true performance dictates. However, this effect was not found in the structured free recall condition on either of our dependent measures. Thus, the impact of gender stereotypes on the accuracy of performance evaluations was diminished by a structured free recall intervention. These results also suggest that the structured free recall intervention is successful at reducing internal performance expectations (i.e., gender-based stereotypes) as well as the external performance expectations (i.e., performance cues) found in previous research.

Limitations, Future Directions, and Conclusions

Several limitations to the present study should be mentioned. First, the participants in the study were only evaluating a set of written vignettes that contained limited information. Thus, given the lack of cognitive complexity in the stimulus vignettes, generalizing to actual performance appraisals may be problematic. The criticisms of laboratory research methods could at least be partially overcome if more realistic business environments were duplicated in the laboratory. One methodology that has been utilized in the interest of enhancing external validity of laboratory studies is the use of videotaped performances in lieu of the paper person approach (used in the present study). Further, the use of videotapes has emerged as the preferred methodology because it is assumed to more closely represent "real" information (Woehr & Lance, 1991). Therefore, future researchers should attempt to assess the relationship between rater stereotype and accuracy of ratings using more realistic stimulus materials.

Second, the success of this intervention only generalizes to situations in which the rater has not had much experience with the ratee and has only had a short amount of time to study the ratee's performance (i.e., an interview or assessment center). In this study the raters had only 5 min to study the vignettes. Given this, we do not know if the intervention would be successful in performance appraisal scenarios that deal with longer time frames (e.g., 6-month performance appraisal reviews). Therefore, future researchers should test structured free recall interventions in appraisal situations with different time frames and with different amounts of information provided. However, these results may generalize to situations where evaluations are made based on limited information, such as the personnel selection interview.

Finally, the scale used to assess stereotypes of women allowed us to investigate whether or not a pro-male bias exists, but did not allow us to investigate if pro-female biases influence ratings. Specifically, the participants who received a low score on the WAPS had a traditional stereotype of women as professors (i.e., associated women with ineffective performance), whereas those with high scores did not associate women professors with low performance. However, a high score on the WAPS does not necessarily indicate a pro-female bias but perhaps just a gender-neutral position on the part of the participant. Evidence of pro-female bias has been found in other

studies. Specifically, when they examined the main effect of professor gender, Rinehart and Young (1996) found a pro-female difference on two factors of professionalism and instruction. Furnham and Duignan (1989) found that individuals with feminist attitudes tended to recall significantly more pro-female and less pro-male information. A close examination of Figs. 1 and 2 suggests that a pro-female bias may be functioning in the present study. That is, as the regression line for male ratees moves across the WAPS scores from low to high, some individuals who do not believe strongly in the traditional stereotype of women are less accurate than those with traditional stereotypes. Thus, this may be an indication of a pro-female bias, however, it is impossible to examine this theory directly due to the nature of the WAPS scale. It is important to note that when looking at the plotted regression lines, the structured free recall also seems to diminish the impact of the pro-female bias. Thus, it may be interesting for future researchers to examine the potential impact of a pro-female bias on the accuracy of evaluations of men. However, one must keep in mind that the pro-female bias is probably far less prevalent than the pro-male bias. Although there have been some stories in the media on discrimination against men and some cases of reverse discrimination have been won in the U.S. court system (e.g., Regents of the University of California v. Bakke, 1978), there is little objective evidence that men are disadvantaged relative to women (Gutek, Cohen, & Tsui, 1996).

In summary, the results of the present study suggest that women who are evaluated by raters who hold traditional stereotypes of women will be at a disadvantage. However, support for an intervention that successfully reduces the bias was also obtained. This is important, as the intervention is one that can easily and cost-effectively be applied in the workplace. Therefore, a structured free recall intervention could be a useful tool for improving the accuracy of performance ratings in the field that involve low information situations (e.g., interviews). Given this, we believe that the present study has made two important contributions to existing performance appraisal research. First, it extends previous research on gender bias by illustrating the influence of gender stereotypes on the accuracy of evaluations of women. Second, it provides support for an intervention that successfully reduces the bias. These are both important advances in understanding the nature of the gender bias and increasing the fairness of evaluations of women.

APPENDIX

Sample Vignette and True Scores

You should now study the performance of Professor 1. The following statements describe the way he performed during the semester. Remember that this information is based upon actual student descriptions. Please read the entire list of statements closely before filling out any part of the evaluation form.

- His workload was so heavy that only 1 out of 25 passed.
- He assigned general problems in class, then gave specific problems on tests.
- He would sometimes get so involved in the subject matter that he would forget to stop lecturing when the class period was over.
- He passed out a mimeographed sheet giving his office hours and office telephone number.
- He tests materials that were not covered in class.
- He offered extra help sessions at night.
- He assigned two papers a week, seven outside books, a textbook, and classroom work for a 2-hr course.
- He has difficulty explaining things simply enough for his students to understand.
- He has a bad accent and is hard to understand.
- He described his own fascination with the material that he was covering.

True scores. The true scores for the above vignette were 8.9 (Relationship With Students), 2.8 (Ability to Present Material), 7.8 (Interest in Course Material), 1.1 (Reasonableness of Workload), and 3.0 (Fairness of Testing and Grading).

REFERENCES

Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin*, *86,* 736–765.

Baltes, B. B., & Parker, C. P. (2000a). Clarifying the role of memory in the performance cue effect. *Journal of Business and Psychology*, *15,* 229–246.

Baltes, B. B., & Parker, C. P. (2000b). Reducing the effects of performance expectations on behavioral ratings. *Organizational Behavior and Human Decision Processes*, *82,* 237–267.

Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review*, *88,* 354–364.

Bodenhausen, G. B., & Macrae, C. M. (1996). The self-regulation of intergroup perception: Mechanisms and consequences of stereotype suppression. In C. M. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 227–253). New York: Guilford.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42,* 116–131.

Cardy, R. L., & Kehoe, J. F. (1984). Rater selective attention ability and appraisal effectiveness: The effect of cognitive style on the accuracy of differentiation among ratees. *Journal of Applied Psychology*, *69,* 589–594.

Cleveland, J. N., Stockdale, M., & Murphy, K. R. (2000). *Women and men in organizations: Sex and gender issues at work*. Mahwah, NJ: Erlbaum.

Cronbach, J. L. (1955). Processes affecting scores on understanding of others and "assumed similarity." *Psychological Bulletin*, *52,* 177–193.

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

Deaux, K., & Taynor, J. (1973). Evaluation of male and female ability: Bias works two ways. *Psychological Reports*, *32,* 261–262.

Del Boca, F. K., & Ashmore, R. D. (1980). Sex stereotypes and implicit personality theory. II. A trait-inference approach to the assessment of sex stereotypes. *Sex Roles*, *6,* 519–535.

Del Boca, F. K., Ashmore, R. D., & McManus, M. A. (1986). Gender-related attitudes. In R. D. Ashmore & F. K. Del Boca (Eds.), *The social psychology of female–male relations: A critical analysis of central concepts* (pp. 121–163). Orlando, FL: Academic Press.

DeNisi, A., Cafferty, T., & Meglino, B. (1984). A cognitive view of the performance appraisal process: A model and some research propositions. *Organizational Behavior and Human Decision Processes*, *33,* 360–396.

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: A laboratory and field study. *Journal of Applied Psychology*, *73,* 551–558.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, *73,* 421–435.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Fiske, S. T. (1991). Social science research on trial: Use of sex stereotyping research in Price Waterhouse v. Hopkins. *American Psychologist*, *46,* 1049–1060.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In. D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 357–411). Cambridge, MA: McGraw-Hill.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum model of impression formation: From category-based to individuating processes as a function of information, motivation, and attention. In M. P. Zanna (Ed.), *Advances in experimental psychology* (Vol. 23, pp. 1–108). San Diego, CA: Academic Press.

Furnham, A., & Duignan, S. (1989). The selective recall of attitude consistent information: A study concerning sex differences. *Psychologia*, *32,* 112–119.

Glick, P., Zion, C., & Neslon, C. (1988). What mediates sex discrimination in hiring decisions? *Journal of Personality and Social Psychology*, *55,* 178–186.

Goldberg, P. A. (1968). Are women prejudiced against women? *Transactions*, *5,* 28–30.

Gordon, M. E., Slade, L. A., & Schmitt, N. (1986). The "Science of the Sophomore" revisited: From conjecture to empiricism. *Academy of Management Review*, *11,* 191–207.

Gunderson, D. E., Tinsley, D. B., & Terpstra, D. E. (1996). Empirical assessment of impression management bias: The potential for performance appraisal error. *Journal of Social Behavior and Personality*, *11,* 57–76.

Gutek, B. A., Cohen, A., & Tsui, A. (1996). Reactions to perceived sex discrimination. *Human Relations*, *49,* 791–813.

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

Jako, R. A., & Murphy, K. R. (1990). Distributional ratings, judgement decomposition, and their impact on interrater agreement and rating accuracy. *Journal of Applied Psychology*, *75,* 500–505.

Kalin, R., & Hodgins, D. C. (1984). Sex bias in judgments of occupation suitability. *Canadian Journal of Behavioral Science*, *16,* 311–325.

Locksley, A., Borgida, E., Brekke, N., & Hepburn, C. (1980). Sex stereotypes and social judgment. *Journal of Personality and Social Psychology*, *39,* 821–831.

Locksley, A., Hepburn, C., & Ortiz, V. (1982). Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology*, *18,* 23–42.

Macdonald, A. P. (1970). Revised scale for ambiguity tolerance: Reliability and validity. *Psychological Reports*, *26,* 791–798.

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, *66,* 37–47.

Martell, R. F. (1996). What mediates gender bias in work behavior ratings? *Sex Roles, 35,* 153–169.

Martell, R. F., & Guzzo, R. A. (1991). The dynamics of implicit theories of work group performance: When and how do they operate? *Organizational Behavior and Human Decision Processes*, *50,* 51–74.

Martell, R. F., & Willis, C. E. (1993). Effects of observer's performance expectations on behavior ratings of work groups: Memory response bias? *Organizational Behavior and Human Decision Processes*, *56,* 91–109.

Martinko, M., & Gardner, W. (1983). A methodological review of sex-related access discrimination problems. *Sex Roles*, *18,* 23–42.

Maurer, T. J., & Taylor, M. A. (1994). Is sex by itself enough? An exploration of gender bias issues in performance appraisal. *Organizational Behavior and Human Decision Processes*, *60,* 231–251.

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

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology*, *67,* 320–325.

Nelson, T. E., Acker, M., & Manis, M. (1996). Irrepressible stereotypes. *Journal of Experimental Social Psychology*, *32,* 13–38.

Nelson, T. E., Biernat, M. R., & Manis, M. (1990). Everyday base rates (sex stereotypes): Potent and resilient. *Journal of Personality and Social Psychology*, *59,* 664–675.

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

Pazy, A. (1986). The persistence of pro-male bias despite identical information regarding causes of success. *Organizational Behavior and Human Decision Processes*, *38,* 366–377.

Peters, L. H., Terborg, J. R., & Taynor, J. (1974). Women as managers scale: A measure of attitudes toward women in managerial positions. *Journal Supplement Abstract Service Catalog of Selected Documents in Psychology*, *4,* 27.

Pratto, F., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology*, *27,* 26–47.

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

Regents of the University of California v. Bakke. (1978). *U.S. Law Weekly*, *49,* 4896.

Rinehart, J. S., & Young, I. P. (1996). Effects of teacher gender and principal gender on the ratings of teacher performance. *Journal of Personnel Evaluation in Education*, *10,* 313–323.

Robbins, T. L., & DiNisi, A. S. (1993). Moderators of sex bias in the performance appraisal process: A cognitive analysis. *Journal of Management*, *19,* 113–126.

Sauser, W. I., Evans, K. L., & Champion, C. H. (1979). *Two hundred and fifty scaled incidents of college classroom behavior.* Paper presented at the Southeastern Psychological Association annual conference, New Orleans.

Seta, J. J., & Seta, C. E. (1993). Stereotypes and the generation of compensatory and noncompensatory expectancies of group members. *Personality and Social Psychology Bulletin*, *19,* 722–731.

Shaw, E. A. (1972). Differential impact of negative stereotyping in employee selection. *Personnel Psychology*, *25,* 333–338.

Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology*, *19,* 174–197.

Snyder, M., Campbell, B. H., & Preston, E. (1982). Testing hypotheses about human nature: Assessing the accuracy of social stereotypes. *Social Cognition*, *1,* 256–272.

Snyder, M., & Gangestad, S. (1986). On the nature of self-monitoring: Matters of assessment, Matters of validity. *Journal of Personality and Social Psychology*, *51,* 125–139.

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

Swim, J. K., & Sanna, L. J. (1996). He's skilled, she's lucky: A meta-analysis of observers' attributions for women's and men's successes and failures. *Personality and Social Psychology Bulletin*, *22,* 507–519.

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

Watson, D., Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scale. *Journal of Personality and Social Psychology*, *54,* 1063–1070.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67,* 1049–1062.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, *101,* 34–52.

Woehr, D. J., & Lance, C. E. (1991). Paper people versus direct observation: An empirical examination of laboratory methodologies. *Journal of Organizational Behavior*, *12,* 387–397.

Yammarino, F. J., & Dubinsky, A. J. (1988). Employee responses: Gender—or job-related differences? *Journal of Vocational Behavior*, *32,* 366–383.