Go to old article view

Personnel Psychology

Early View

Online Version of Record published before inclusion in an issue

Main Section Submission

When Accomplishments Come Back to Haunt You: The Negative Effect of Competence Signals on Women's Performance Evaluations

M. Ena Inesi , Daniel M. Cable

First published:
25 July 2014 Full publication history

DOI:
10.1111/peps.12083 View/save citation

Cited by:
0 articles Check for new citations

score 27

The authors would like to thank Selin Kesebir, Madan Pillutla, Stefan Thau, our action editor, and the anonymous reviewers for providing helpful feedback on this manuscript.

Abstract

This research explores the possibility that the very accomplishments that are critical to success during the hiring process (e.g., educational attainment, promotion history) can lead to a drop in future performance evaluations for women. We theorized that evaluators may see such competence signals as a threat to the traditional gender hierarchy, which leads to a negative bias when evaluating women's on-the-job performance. In Study 1, we examined this hypothesis among commanding officers in the U.S. military, who gave lower performance ratings to female subordinates whose pay grade approached their own. The same was not true for male subordinates. Studies 2, 3a, and 3b experimentally tested the boundary conditions of this effect using two additional competence signals (educational attainment and past career successes) and 2 different populations. Across these studies, we replicated the negative relationship between competence signal strength and performance evaluations for female subordinates but only under conditions in which the evaluator would be particularly likely to experience gender hierarchy threat. Specifically, it emerged when the evaluator was male and high social-dominance oriented and when the female subordinate's objective on-the-job performance was high. Finally, Study 3a demonstrated how organizations can mitigate this negative bias by using objective (rather than subjective) performance evaluations.

Enhanced Article Feedback

When applying for a job, we all put our best foot forward. We fill our resumés with our best accomplishments in the hopes that they will provide a clear signal of our competence and help us obtain job offers. According to signaling theory (Bangerter, Roulin, & König, 2012 ; Spence, 1973), “competence signals” such as educational attainment, job level, and past job accomplishments can be used by organizations to make accurate inferences about applicant ability. Thus, the more competence signals you can provide as a job applicant, the more likely you are to obtain a job offer (Spence, 1973).

But once you are successful in getting hired, how might these same competence signals affect your subsequent performance evaluations? According to theories of social influence, status characteristics, and self-fulfilling prophecies, evaluators may remain more positively disposed toward employees with greater competence signals (Berger, Cohen, & Zelditch, 1972 ; Cialdini, 1993 ; Rosenthal & Jacobson, 1966). In other words, because we often perceive and construct the social world that we expect to see, these theories would imply that people who were hired with stronger competence signals will receive better future evaluations than those with weaker competence signals—even controlling for objective on-the-job performance.

In this research, however, we propose and test the counterintuitive prediction that the opposite dynamic can occur for female subordinates. We propose that women who were hired with stronger competence signals may receive *worse* evaluations than women with weaker competence signals, even if the work being evaluated is identical. For example, a woman who achieved a higher education degree may receive a lower performance evaluation than a woman who did not achieve this degree, all else being equal. We build our theory on four main points: first, competence signals are associated with status in the workplace. Second, social status is incongruent with the lower status role that women hold in the traditional gender hierarchy (Rudman, Moss-Racusin, Phelan, & Nauts, 2012). Third, an evaluator who is threatened by status incongruence will be biased against women with stronger competence signals. Fourth, the subjectivity inherent in performance evaluations permits bias to enter.

By focusing on the complex relationship between gender, competence signals, and performance evaluations, our research identifies a previously unknown hurdle faced by women in the workplace. Although women are more likely to be hired when they show stronger competence signals (although, notably, they benefit less than men do; Lyness & Heilman, 2006), these same signals can have a detrimental effect on later performance evaluations. This hurdle is particularly problematic because it implies that the very competence signals that help a woman get hired (such as education level and demonstrated promotion history) come back to harm her down the line.

Gender and Career Outcomes

One of the greatest challenges faced by working women is that they are assumed to be less competent than men. Because the stereotype for women—warm and nurturing—is inconsistent with the drivers of career success—agency, assertiveness, and dominance (Eagly & Karau, 2002 ; Heilman, 1983 ; Nieva & Gutek, 1980 ; Schein, 1973 , 1975)—women's competence is consistently undervalued (Dobbins, Cardy, & Truxillo, 1988 ; Lyness & Judiesch, 1999 ; Tosi & Einbinder, 1985). To overcome this initial hurdle and provide evidence of their competence (Hewlett, Peraino, Sherbin, & Sumberg, 2010), many women focus their efforts on objective accomplishments, such as earning a degree from a prestigious university, excelling on standardized tests, or demonstrating a history of successful promotions. Consistent with signaling theory research, we define these accomplishments as *competence signals* .

Competence signals are indeed important to career success because they are used by employers to decide whom to hire. This is because, according to signaling theory (Spence, 1973), hiring situations are characterized by informational asymmetry: Applicants have full information about their abilities, whereas firms cannot observe this quality. Therefore, quality applicants may preemptively pursue accomplishments that signal their superior ability, such as a challenging degree at a top university or high performance levels in a prior job. These are called honest competence signals because the costs of accomplishing them are inversely related to the quality of the applicant (Connelly, Certo, Ireland, & Reutze, 2011). In other words, the cost of attaining honest competence signals is higher for less qualified individuals, and therefore they are less likely to pursue them. Thus, by listing competence signals on a resumé or job application, applicants increase their chance of being hired.

In this research, however, our focus is not on the direct effect of competence signals on hiring but on the downstream consequences of signals for later performance evaluations. Unlike hiring decisions, performance evaluation situations are characterized by symmetrical information. Subordinate performance is observable to evaluators, meaning that performance evaluations should be based on job performance itself not historic competence signals. As such, such signals from the past should not affect future performance evaluations.

At the same time, we know that many variables unrelated to actual performance can creep into and affect supervisors' performance evaluations. Research has shown that performance ratings are subject to implicit cognitive biases (DeNisi & Williams, 1988) and are affected by distal factors not directly related to an employee's performance such as interpersonal attraction (Cardy & Dobbins, 1986 ; Dipboye, 1985 ; Ilgen & Feldman, 1983), beliefs about an employee's personality (Krzystofiak, Cardy, & Newman, 1988), and demographic information such as gender, age, and race (Berger et al., 1972 ; Foschi, 1992). As such, the ambiguity inherent in performance evaluations creates an opportunity for bias. Even though competence signals may be orthogonal to a subordinate's on-the-job performance, logically they can affect performance evaluations by biasing evaluators toward or against the subordinate.

Signals, Status, and Performance Evaluations

To the extent that performance evaluations are prone to bias, we might predict that stronger competence signals would exert a positive bias. That is, strong competence signals—which by definition are signals of competence—might lead an evaluator to give a subordinate the benefit of the doubt under conditions of uncertainty or subjectivity. For example, a core tenet of expectation states theory (Berger, Rosenholtz, & Zelditch, 1980) is that a personal characteristic that is associated with more positive expectations in one domain can also generate more generalized positive expectations. These generalized expectations affect a host of outcomes, including more positive evaluations of that individual (Berger, Conner, & McKeown, 1969 ; Berger et al., 1980). Relatedly, there is considerable evidence that recruiters' post-interview evaluations of job applicants are unduly biased by knowledge of the applicants' competence signals (Cable & Gilovich, 1998 ; Dipboye, 1982 ; Dougherty, Turban, & Callendei, 1994). In other words, evaluators tend to overweight competence signals and underweight information gathered during the interview, such as ability to answer questions and fit with the organization.

At the same time, competence signals connote status. According to status characteristics theory, status is granted to those group members for whom positive expectations are held about their potential contribution to a group task (Berger et al., 1972). In one study, for example, participants with greater educational attainment were given greater responsibility in a group task compared to those with less educational attainment (Zeller & Warnecke, 1973). In another study, individuals with higher military rank were more influential in a group than those with lower rank (Driskell, 1986 , as cited in Driskell & Mullen, 1990).

In this research, however, we propose that the status associated with competence signals conflicts with the (low) status of a woman's gender, leading to a counterintuitive negative effect of competence signals on performance evaluations. Across all cultures, men hold higher-status positions compared to women (Buss, 1989 ; Connell, 1987), providing them greater power, wealth, and access to desired occupations (Umphress, Simmons, Boswell, & del Carmen Triana, 2008). When a woman gains status through her competence signals, this creates a sort of oxymoron: a high-status woman. Further, this “status incongruence” (Rudman et al., 2012) calls into question the legitimacy of the gender hierarchy because it blurs the distinction between the sexes (i.e., distinctiveness threat, Branscombe, Ellemers, Spears, & Doosje, 1999). Therefore, evaluators may feel threatened by a woman with strong competence signals. Although the experience of threat may be particularly strong for those most invested in the extant hierarchy, even those who benefit less may experience some degree of threat. For example, Berdahl (2007) writes that, consistent with theories of system justification (Jost & Banaji, 1994), women can hold beliefs that reinforce male dominance and experience threat when it is challenged.

Research shows that people—often unconsciously—act against perceived threats in order to reduce them (e.g., Brehm, 1966). This suggests that evaluators might show a negative bias against a female subordinate with stronger (vs. weaker) competence signals. Some empirical work supports this dynamic. In one study, for example, participants learned that a confederate scored higher or lower on a leadership aptitude test than they did (i.e. the confederate showed evidence of higher or lower status than the participant). A high-scoring female confederate was more likely to be sabotaged than either a high-scoring male confederate or a low-scoring (male or female) confederate (Rudman et al., 2012 , Study 5).

In summary, women seek to accrue competence signals to increase their likelihood of getting hired. However, these same competence signals may negatively bias future evaluators who are threatened by the associated status incongruence. Therefore:

Hypothesis 1 **:**

All else being equal, an evaluator will provide a lower performance evaluation to a female subordinate with stronger competence signals compared to a female subordinate with weaker competence signals.

So far, we have focused on the negative relationship between competence signals and current performance evaluations for female subordinates. Our prediction is predicated upon these women representing a threat to the gender hierarchy. This logic suggests that the same negative relationship would not exist for male subordinates, whose competence signals do not threaten the gender hierarchy. It is possible that male subordinates would benefit from the positive expectancy associated with competence signals and receive more positive evaluations as a result of stronger competence signals. It is also possible that their evaluations could be unaffected by competence signals, which technically do not bear on the performance being evaluated. Although the specific nature of the within-male contrast is beyond the scope of the status incongruence hypothesis, it does suggest an interaction between subordinate gender and competence signal strength. In sum, we predict that:

Hypothesis 2 **:**

Subordinate gender moderates the effect of competence signal strength on performance evaluations, such that stronger competence signals lead to worse performance evaluations for female subordinates but not for male subordinates.

Study 1

In Study 1, we studied a branch of the U.S. military where evaluators were commanding officers from the combat-oriented part of the military. The U.S. Armed Forces is one of the world's largest employers, with 2,266,883 troops serving on active duty, in the National Guard, in the Air National Guard or in the reserves as of March 31, 2010 (NPR, 2011). In our study, each commanding officer provided a performance evaluation for his or her legal advisor, who is a professional with a law degree, deployed in the field to serve a commanding officer by providing legal services, information, and solutions. The commanding officers are part of the tactical branch of the military, whereas legal advisors are part of the legal branch of the military.

We operationalized competence signal through individuals' pay grades, a clear signal of competence in the military. Pay grade is an especially salient competence signal in the military because it is evident on employees' uniforms and in the manner of addressing coworkers (e.g., Lieutenant, Ensign). Despite the legal branch of the military being separate from the central, combat-oriented hierarchy, both rely on the same pay-grade system as a means of indicating career accomplishment. Pay grades generally range from 1 to 10, with higher numbers indicating higher status. In our sample, evaluators (commanding officers) are promoted into higher pay grades for tactical experience and successes, whereas their legal advisor subordinates are promoted into higher pay grades for legal experience and successes. Further, because pay grade implies income level and occupational attainment in the military, it also reflects socioeconomic status (Oakes & Rossi, 2003), which is the central measure of societal status in the United States (Haug & Sussman, 1971).

One interesting feature of our evaluator–subordinate pairs is that their pay grades are significantly but not perfectly correlated ($r = .539$, $p < .001$). An evaluator has either a higher or equal pay grade compared to the subordinate. At the same time, the strong positive correlation demonstrates that as an evaluator's pay grade rises, so does his subordinate's, creating a sort of moving window of relative competence signals. This unique feature of our data may mute the experience of gender hierarchy threat because, as past research has shown, system threats are triggered by situations that are more personally relevant (for example, see Lowery, Knowles, & Unzueta, 2007 ; Rudman & Fairchild, 2004 ; Unzueta & Lowery, 2008). With this in mind, we computed a measure of pay-grade proximity, which we elaborate on below. We predicted that a commanding officer's performance evaluation of a female legal advisor would drop as her pay grade approached his own. We further predicted that this would not be the case for male legal advisors.

Method

Participants

We investigated evaluator–subordinate pairs. Each evaluator had been assigned a legal advisor and was responsible for evaluating his or her job performance. We sent our survey to 267 commanding officers (evaluators), and 193 people responded (72%). In our study, each legal advisor's pay grade is either lower than or equal to his or her commanding officer's pay grade.

For the commanding officers, the average age was 50.12 years (*SD* = 4.45), and they were 96.8% male. Their average pay grade was 6.76 (*SD* = 1.18, Min = 5, Max = 10), reflecting their senior status in the military. Of those for whom we could obtain race data (96.9% of respondents), they were 93.6% White, 1.6% African American, 1.1% Hispanic/Latin American, 2.1% Asian or Pacific Islander, 1.1% American Indian or Alaskan native, and .5% Other.

For the legal advisors, the average age was 39.59 years (*SD* = 6.50), and they were 62.7% male. Their average pay grade was 4.34 (*SD* = 1.07, Min = 3, Max = 6), and their average tenure in the military was 14.83 years (*SD* = 6.16). For these subordinates, age and tenure in the military were highly correlated, ($r = .885$, $p < .001$).[1] Of those for whom we could obtain race information (80.8% of subordinates), they were 92.3% White, 2.6% African American, 1.3% Hispanic/Latin American, and 3.8% Asian or Pacific Islander.

Design and procedure

We gathered the gender and the pay grades of the commanding officers and their subordinates from the military's records. To measure performance evaluations, we worked with the top leader of the legal organization to develop three items that would best reflect the key dimensions of an actual appraisal of a legal advisor's performance. We then emailed an Internet survey link to the commanding officers. The items asked to what extent commanding officers agreed with the following statements: "I am completely satisfied with [this person's] performance," "[This person] is part of my inner circle of confidence," and "[This person] is absolutely critical to my decision making" (Cronbach's α = .94). Possible responses ranged from 1 = *strongly disagree to 5 = strongly Agree* .

We computed pay-grade proximity by subtracting the supervisor's pay grade from his subordinate's, such that a larger difference score indicates greater proximity ($\mathrm{Min}_{\mathrm{prox}}$ = –6, M_{axprox} = 0, M_{prox} = –2.42, SD_{prox} = 1.09). Despite their simplicity and prevalence in the organizational behavior literature, difference scores make assumptions about the nature of the relationships under investigation (Edwards, **1995**). Accordingly, before examining the effects of our difference score we tested these assumptions to ensure its appropriateness. We first tested for an adequate range of discrepancy using the method outlined by Fleenor and colleagues (**1996**). We found reasonable dispersion, with 61.6% of our dyads characterized by some degree of pay-grade discrepancy, whereas the remaining 38.4% had no pay-grade discrepancy. Next, difference scores assume a linear relationship between the predictor variables and the outcome variable, and thus we tested whether the two variables comprising the difference score exerted a linear or a quadratic effect on the outcome measure. To do this, we entered the subordinate and the supervisor pay-grade variables and the subordinate gender variable, plus the relevant interaction terms in Step 1. We then entered the relevant quadratic terms in Step 2. We found that the increase in R^2 after Step 1 was significant p = .009, whereas the additional benefit of adding the quadratic terms in Step 2 was not significant, p = .680. These results support the assumption that the two variables in our difference score exert linear effects on the dependent measure. Finally, difference scores assume that the two subvariables exert an equal and opposite effect on the dependent variable (Edwards & Parry, **1993**). The results of the first model described above confirmed that the subordinate pay grade × subordinate gender effect (β = –.19, t = –1.75, p = .082) was approximately equal and opposite to the supervisor pay grade × subordinate gender effect (β = .20, t = 1.79, p = .076). Together, these results support the use of a difference score.

Results

We were unable to test the effect of evaluator gender due to the small number of female commanding officers, and we have therefore removed female supervisors from the analysis.[2] In addition, we could not identify the subordinate's gender or pay grade for two responses and thus have dropped these from the dataset. This left 186 supervisor–subordinate pairs that had been working together an average of 15.58 months (SD = 8.13).[3] Table 1 presents the correlation matrix.

Table 1. Correlation Table for Study 1

Measure	***Mean***	***SD***	**1**	**2**	**3**
1. Evaluator pay grade	6.77	1.19			
2. Subordinate pay grade	4.34	1.07	.539**		
			**	**	

3. Pay-grade proximity	−2.43	1.09	−.562	.394	
4. Subordinate tenure in military (years)	14.83	6.16	−.024	.024	.050

**Correlation is significant at the .01 level (2-tailed).

We tested our prediction using hierarchical multiple regression (Cohen, Cohen, West, & Aiken, 2003). We regressed the performance evaluation provided by the evaluator onto subordinate gender (0 = *male*, 1 = *female*), pay-grade proximity (mean-centered) and their interaction.[4] Subordinate tenure in military was also included as a covariate to rule out the possibility that any effect that emerges could be attributed to the amount of time the subordinate had spent in the military. Step 1 revealed a nonsignificant effect of tenure. Step 2 revealed a main effect of subordinate gender, (β = –.22), *t* (179) = –3.10, *p* = .002, and a marginal effect of pay-grade proximity, (β = –.14), *t* (179) = –1.90, *p* = .059. In Step 3, the two-way subordinate gender × pay-grade proximity interaction was significant (β = –.19), *t* (178) = –1.99, *p* = .049. See Table 2 for step-wise regression results.

Table 2. Regression Results From Study 1

	Step 1			Step 2			Step 3		
Variable	***b***		**95% CI**	***b***		**95% CI**	***b***		**95% CI**
Subordinate years in military[a]	–.01		[−.03, .01]	−.01		[−.03, .01]	−.01		[−.03, .01]
ΔR^2		<.01							
$F(\Delta R^2)$		.41							
Subordinate gender[b]				−.38**		[−.62, −.14]	−.38**		[−.62, −.14]
Pay-grade proximity[c]				−.10^{+}		[−.21, .004]	−.02		[−.15, .12]
ΔR^2					.07				
$F(\Delta R^2)$					6.42**				
Subordinate gender × Pay-grade proximity							−.21*		[−.43, −.001]
ΔR^2								.02	
$F(\Delta R^2)$								3.94*	
R^2		<.01			.07**			.08**	

Note

Stepwise regression model. *F* -test indicates model fit increase. CI = confidence interval.

a Mean-centered.

b 0 = male, 1 = female.

c Subordinate pay grade – supervisor pay grade. Larger numbers indicate subordinate pay-grade approaches supervisor's.

+ <.10, * *p* < .05. ** *p* < .01. *** *p* < .001.

As shown in Figure 1, decomposition of this interaction revealed that for male subordinates, there was no effect of pay-grade proximity on performance evaluation, (β = –.02), *t* (178) = –.21, *p* = .831. However for female subordinates, as her pay grade approached that of her supervisor, she received a lower performance evaluation, (β = –.31), *t* (178) = –2.75, *p* = .007.

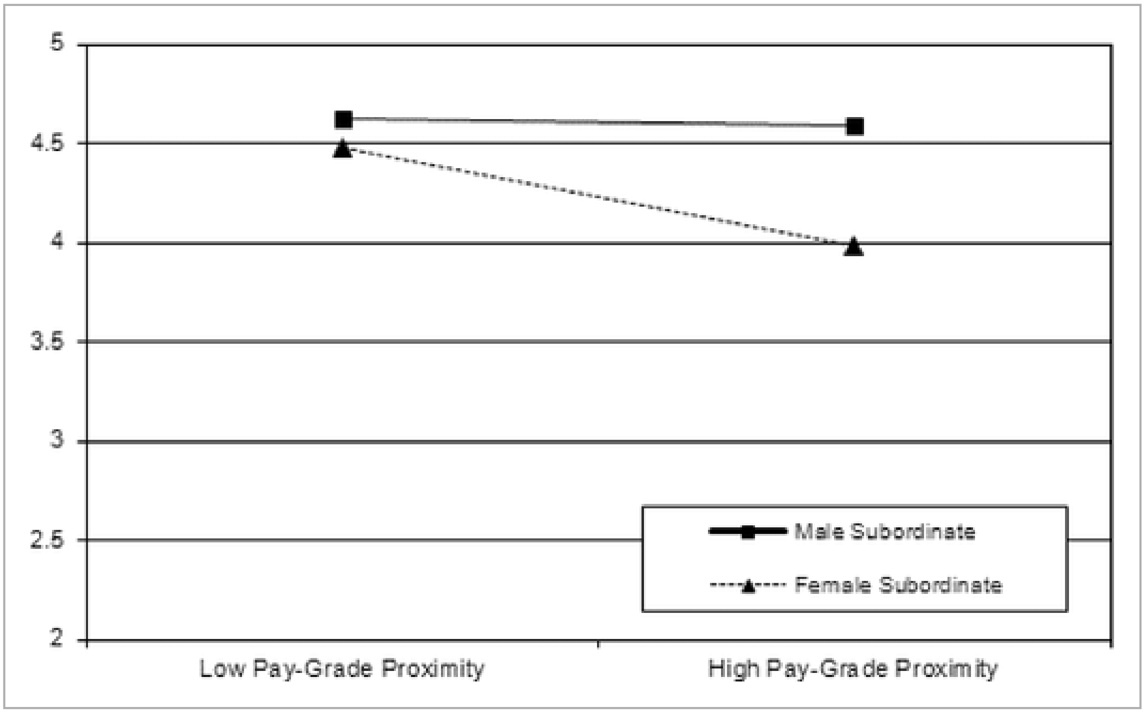

Figure 1.

Open in figure viewer

The Effect of Subordinate Competence Signal (Pay-Grade Proximity) on Performance Evaluations by Male Evaluators, Shown Under Different Conditions of Subordinate Gender, Controlling for Subordinate Years in Military (Study 1).

Discussion

Results from Study 1 provide some initial evidence consistent with Hypothesis 1: Evaluators provided lower performance evaluations for female subordinates whose competence signals approached their own. This did not occur for male subordinates, leading to the interaction predicted by Hypothesis 2. However, a clear limitation of Study 1 is that we were unable to control for the subordinate's actual performance (i.e., on-the-job accomplishment). Therefore, it is possible that women's performance objectively dropped as their pay grade approached that of their supervisors. This could have occurred either because the supervisor treated them differently, eliciting worse behavior (i.e., stereotype confirmation, Word, Zanna, & Cooper, 1974), or because they were simply worse at the job.

A second limitation of Study 1 is that the competence signal (pay-grade) is likely not independent of the performance being evaluated. Because pay grade is a reward for ability, higher pay-grade subordinates are probably given more difficult assignments and provide better quality output. If anything, this should lead to more positive performance evaluations for both male and female subordinates with higher pay grades. However, it is important to demonstrate the predicted effect in a context in which the competence signal is independent of the performance being evaluated.

Third, although our results provide evidence consistent with Hypotheses 1 and 2, the setting was characterized by a number of features that may have made evaluators feel particularly threatened by women's status and therefore made our predicted effects especially likely to emerge. Specifically, the evaluators were virtually all men and, as senior members of the military, were likely high social dominance-oriented individuals (Nicol, Charbonneau, & Boies, 2007). Moreover, the subordinate performance being evaluated was viewed as extremely strong (average evaluations were 4.5 out of 5).

In the next section, we develop theory as to why these variables may exacerbate gender hierarchy threat in evaluators. Then in Studies 2, 3a, and 3b, we experimentally manipulate these variables to better understand what features of the environment need to be in place for the predicted bias to emerge. In addition to providing important information that can help practitioners understand the conditions leading to bias, these moderators also help provide evidence of our predicted mechanism (Spencer, Zanna, & Fong, 2005), which we elaborate on below.

Boundary Conditions of Threat Response

Our basic premise is that the negative relationship between competence signals and performance evaluations for female subordinates emerges because evaluators experience them as a threat to the traditional gender hierarchy. Theoretically, certain characteristics of evaluators and subordinates should make evaluators particularly sensitive to this threat. Specifically, evaluators who are most invested in maintaining the existing gender hierarchy should be most prone to a negative bias because they will lose more from a change. Because men have more to lose than women from a shift in the traditional gender hierarchy (Dall’Ara & Maass, 1999), male evaluators should be more likely than female evaluators to be negatively biased against female subordinates with stronger competence signals (see also Berdahl, 2007 , for a related argument). Consistent with this claim, a series of studies has found that men are more likely than women to react negatively to gender stereotype violators (Carli, LaFleur, & Loeber, 1995 ; Rudman, 1998), and male respondents are less likely than female respondents to see an overlap between stereotypes of women and stereotypes of managers (Duehr & Bono, 2006).

We also predict more negative bias from evaluators who support hierarchy maintenance in general (Berdahl, 2007 ; Dobbins et al., 1988 ; Rudman et al., 2012). According to social dominance theory, certain individuals have higher social dominance orientation (SDO) and are more likely to “support group-based hierarchy and the domination of ‘inferior’ groups by ‘superior’ groups” (Sidanius & Pratto, 1999 , p. 48). A person's SDO has been linked to discriminatory attitudes and behaviors (Altemeyer, 1998). For example, high SDO participants had more negative attitudes toward Black employees (Aquino, Stewart, & Reed, 2005) and were less likely to select a female or a Black individual to join their group (Umphress et al., 2008). Because the existing gender hierarchy places men above women, high SDO individuals should be more likely to support the current gender hierarchy and be biased against women who challenge it.

Finally, we examine the effect of a subordinate's current job performance. If in addition to demonstrating strong competence signals from past accomplishments, a woman's objective, current on-the-job performance is very strong, this should represent an even greater threat to the traditional gender hierarchy. For example, Rudman and colleagues demonstrated that when a woman shows high on-the-job performance, it evokes a threat response in subordinates (e.g., Rudman et al., 2012 , Study 5).

Evidence for Mechanism

A key contribution of our research is to demonstrate that the negative relationship between competence signals and performance evaluations for female subordinates occurs *because* evaluators experience threat to the gender hierarchy. In other words, threat to the gender hierarchy is our theoretical mechanism. Unfortunately, assessing threat poses a classic measurement problem because often people are not consciously aware of it, and even if they are they may be loath to admit it (e.g., Scheepers & Ellemers, 2005). As discussed by Blascovich and colleagues (2001), self-reports of threat are open to reactivity and defensive reactions, leading those who are most threatened to indicate this to the least extent. Therefore, evidence of process by directly measuring threat is impractical.

However, it is possible to gather evidence about the threat process by isolating and testing key moderator variables (moderation-of-process design; Spencer et al., 2005), such as those described in the previous section. We have theorized that evaluator gender, evaluator SDO, and subordinate on-the-job performance will increase the gender hierarchy threat experienced by an evaluator. If we manipulate these variables experimentally, and show that the effect increases in strength when these variables are present, but decreases when they are absent, then we can infer that gender hierarchy threat is the mechanism underlying our effect (see Kay et al., 2009 ; Rudman et al., 2012 for recent examples).

Evidence of moderation by these variables would also help rule out alternative plausible mechanisms. For example, it is possible that women with strong competence signals receive worse performance evaluations than women with weak competence signals because evaluators see competence signals as prescriptive norm violations that require sanctioning rather than because they experience threat. Prior work has demonstrated that women who act in agentic ways elicit negative reactions and disapproval from others because they are seen to have behaved in ways that are prescriptively proscribed (e.g., Heilman & Okimoto, 2007). Because prescriptive norms are group based and presumably known by all members of that group, the associated negative responses should be elicited in all group members. Indeed, Okimoto and Brescoll (2010) write that “prescriptive expectations are pervasive [and] typically endorsed by both men and women” (p. 24). Thus, if our effect were driven by negative reactions to a gender norm violation, then it would not vary by factors that we have predicted will give rise to greater threat, such as gender or SDO of the evaluator. However, if our effect is moderated by such evaluator characteristics, then this not only supports gender hierarchy threat but also rules out norm violation as an alternative explanation. In sum, we predict:

Hypothesis 3 **:**

Characteristics of the subordinate–evaluator relationship that heighten gender hierarchy threat—including evaluator gender (male), evaluator social dominance orientation (high), and subordinate on-the-job performance (high)—will strengthen the negative relationship between competence signals and performance evaluations for female subordinates.

Reducing Evaluator Bias

A critical element of the theory we have developed thus far is that the negative bias against women with strong competence signals can emerge on performance evaluations because they are subjective, in that they require interpretation by supervisors. Conceptually, a performance evaluation can be relatively unambiguous and based on objective job results (e.g., booked sales). However, many organizations use approaches to performance evaluation that require substantial interpretation regarding the quality and value of employees (Schmidt & Hunter, 1992). As noted by Nieva and Gutek (1980), “the greater the amount of inference required in the evaluation situation, the more likely it is that evaluation bias will be found.” In considering interventions that managers can use to reduce the potential for bias to emerge, we predicted that the more objective the performance evaluation, the less likely the effect would be to emerge.

Hypothesis 4 **:**

The more subjective the performance evaluation, the more likely the negative relationship between competence signals and performance evaluations for female subordinates is to emerge.

In Studies 2, 3a, and 3b, we test Hypotheses 1–4 and address limitations inherent in the field setting.

Study 2

Study 2 was designed to accomplish three main goals. First, we sought to rule out the possibility that the pattern of results in Study 1 was driven by changes in objective on-the-job performance. To rule this out, we created a procedure that holds constant the subordinate's objective on-the-job accomplishment across our manipulated conditions. Two additional goals of Study 2 were to test the boundary conditions of the effect that emerged in Study 1 and provide evidence of the proposed mechanism driving the effect. Thus, we included both male and female evaluators in Study 2, and we assessed participants' SDO, which is a validated measure of individuals' preference for hierarchy (Pratto, Sidanius, Stallworth, & Malle, 1994). If these variables moderate the effect that emerged in Study 1 and in the predicted direction, then this would illuminate the boundary conditions of Study 1's findings and provide evidence of mechanism (see "Evidence for Mechanism" section above).

Finally, in Study 2 we sought to replicate the evidence consistent with Hypotheses 1 and 2 in a new work context and with a new operationalization of competence signals. Our operationalization of competence signals in Study 2 was educational attainment (either high-school diploma or college degree). Like pay-grade level in Study 1, education level is a signal that is available on people's resumés, is not immediately relevant to current performance, but still may bias an evaluator's responses because it may affect general attributions about the subordinate. To conceptually replicate the relative pay-grade structure in Study 1, we recruited college-educated individuals, such that evaluators would be more educated than a subordinate with a high-school education but would have a similar educational attainment to a subordinate with a college degree.

Method

Participants

Two hundred seventy-one college-educated adults (108 male) participated in an online study. Participants' mean age was 35.38 years (*SD* = 11.26). They were 75.3% White/European American, 8.5% Asian American, 5.5% African American, 3.3% Native American, 3.0% Latin American, and 4.4% Other. Ninety-five percent reported English as their first language, and all reported living in the United States. Fifty-seven percent reported having a full-time job, and 13% reported being a student currently. They were recruited via an online participant pool (Mechanical Turk; see Buhrmester, Kwang, & Gosling, 2011 for a description of this population) and were paid for their participation.

Design and procedure

The experiment was a 2 (subordinate competence signal: weak [high-school education] vs. strong [college education]) × 2 (subordinate gender: male vs. female) × 2 (evaluator gender: male vs. female) between-participants design, with evaluator SDO as a continuous measure.

Participants were asked to act as managers in a company that develops creative ideas to solve their clients' problems. Before working on the task, they were told that they would be working with a subordinate, whom they would evaluate at the end of the study. There were actually no real subordinates; rather we created fake subordinate profiles and fake subordinate input to manipulate the relevant variables and hold constant objective on-the-job performance.

Participants were told that their subordinate was another participant who had previously completed another online study and had allowed us to share demographic information. The cover story for sharing this demographic information was that work groups function more effectively when they know each other (Polzer, Milton, & Swann, 2002), so demographic information would help facilitate that process. This demographic information conveyed our key manipulations. The subordinate's “name” was blocked out (“XXXXXX”), but his/her age (26), gender (either male [*male subordinate*] or female [*female subordinate*]), highest educational attainment (high-school degree [*weak competence signal*] or college degree [*strong competence signal*]), and work status (employed full time) were shared with the participant.

Next, participants were introduced to the project plan for the task, which stipulated that they would (a) read over their client's problem, (b) brainstorm ideas for as long as they wished, (c) request to see their subordinate's ideas when they were ready, and then (d) submit a final idea to the client. The problem was adapted from Johnson and Johnson (**2009**) and described a buyer (i.e., the client) who had purchased 20,000 pipe cleaners for a wholesaler whose warehouse had then burned down. The client had to figure out what to do with the 20,000 pipe cleaners. An image of pipe cleaners was included below this description. All participants were then shown a list of 10 ideas, ostensibly generated by their subordinate, including “Use to clean your cell phone screen” and “Use to make crafts (flowers, fake glasses, hearts for Valentine's Day, etc).” Thus, we held objective performance constant across all conditions. Participants were then given time to develop a single solution and submit it to the client.

After participants submitted their idea to the client, they evaluated their subordinate. We used a multi-item measure of performance evaluation and asked supervisors to answer the following questions: “Please rate your subordinate's performance using the scale below.” (seven-point scale: 1 = *extremely bad* , 4 = *neither bad nor good* , 7 = *extremely good*), “How much do you think your subordinate contributed to the final advice you submitted?” (five-point scale: 1 = *not at all* , 5 = *a great deal*), and “How happy were you with your subordinate's performance?” (five-point scale: 1 = *not at all* , 5 = *extremely*). These items were combined to form a composite measure of performance evaluation (first item transformed to a five-point scale; α = .91).

Before providing demographic information, participants completed the SDO scale (Pratto et al., **1994**). This scale includes 16 items that measure individual differences in preference for hierarchy and group-based domination and discrimination, including such items as “To get ahead in life, it is sometimes necessary to step on other groups” and “Group equality should be our ideal (reverse-scored)” (1 = *strongly disagree* , 7 = *strongly agree*). These items were combined to form the SDO composite (α = .93).[5]

Results

We predicted that there would be a negative relationship between competence signals and performance evaluation for female subordinates (Hypothesis 1) and that this effect would be significantly weaker for male subordinates (Hypothesis 2). Further, we predicted that the negative relationship between competence signals and performance evaluations for female subordinates would be stronger for male evaluators and evaluators with high SDO (Hypothesis 3). We tested these predictions using hierarchical multiple regression (Cohen et al., **2003**). The overall effects of subordinate competence signal (0 = *weak* , 1 = *strong*), subordinate gender (0 = *male* , 1 = *female*), evaluator gender (0 = *male* , 1 = *female*), and evaluator SDO (mean-centered) were entered into the first block, the two-way interaction terms were added to the second block, the three-way interaction terms were added in the third block, and the four-way interaction term in the fourth block.

Step 1 revealed no significant effects. Step 2 revealed a subordinate gender × evaluator gender interaction, ($\beta = –.30$), $t(260) = –2.57$, $p = .011$. Step 3 revealed no significant three-way interactions. In Step 4, the four-way interaction approached significance ($\beta = .44$), $t(255) = 1.94$, $p = .053$. See Table 3 for step-wise regression results. Decomposing this 4-way interaction revealed that the only set of conditions under which the relationship between competence signal strength and performance evaluation approached significance was female subordinates being evaluated by high SDO, male evaluators, ($\beta = –.39$), $t(255) = –1.93$, $p = .055$. All other conditions yielded ps > .15, indicating that the relationship between competence signal strength and performance evaluations for male subordinates was not significant under any condition. See Figure 2 for a depiction of the weak versus strong competence signal slope across each of the experimental conditions. Neither evaluator gender, $F(1, 132) = 3.31$, $p = .071$, $\eta_p^2 = .024$, nor evaluator SDO, ($\beta = .15$), $t(132) = –1.15$, $p = .250$, alone was enough to elicit the negative relationship between competence signal strength and performance evaluation for female subordinates.

Table 3. Regression Results From Study 2

Variable	**Step 1**			**Step 2**		
	b		**95% CI**	***b***		**95% CI**
Subordinate gender[a]	.02		[−.25, .29]	.42		[−.09, .94]
Subordinate past accomplishment (educational attainment)[b]	.12		[−.15, .39]	−.08		[−.58, .41]
Evaluator SDO[c]	−.01		[−.13, .11]	.01		[−.24, .26]
Evaluator gender[d]	.10		[−.18, .37]	.32		[−.14, .79]
▵ R^2		.01				
F (▵ R^2)		.33				
Subordinate gender × Subordinate past accomplishment				.08		[−.46, .61]
Subordinate gender × Evaluator SDO				−.04		[−.29, .22]
Subordinate gender × Evaluator gender				−.72*		[−1.28, −.17]
Subordinate past accomplishment × Evaluator SDO				−.03		[−.27, .22]

Subordinate past accomplishment × Evaluator gender		.24		[−.31, .79]
Evaluator SDO × Evaluator gender		−.01		[−.26, .25]
▵ R^{2}			.03	
F (▵ R^{2})			1.22	
Subordinate gender × Subordinate past accomplishment × Evaluator SDO				
Subordinate gender × Subordinate past accomplishment × Evaluator gender				
Subordinate gender × Evaluator SDO × Evaluator gender				
Subordinate past accomplishment × Evaluator SDO × Evaluator gender				
▵ R^{2}				
F (▵ R^{2})				
Subordinate gender × Subordinate past accomplishment × Evaluator SDO × Evaluator gender				
▵ R^{2}				
F (▵ R^{2})				
R^{2}	.01		.03	

Note

Stepwise regression model. F-test indicates model fit increase. CI = confidence interval.

a 0 = male, 1 = female.

b 0 = high school education, 1 = college education.

c Mean-centered.

d 0 = male, 1 = female.

+ $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

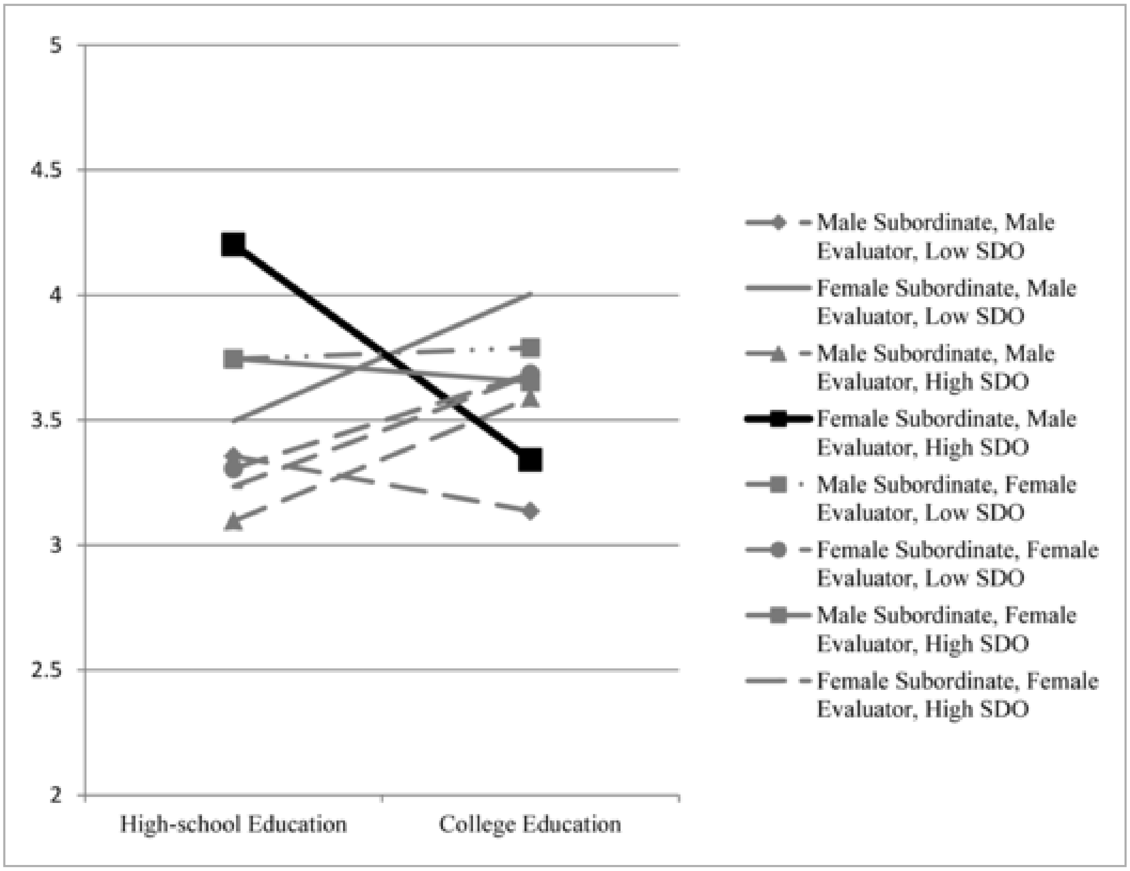

Figure 2.

Open in figure viewer

The Effect of Subordinate Competence Signal (Educational Attainment) on Performance Evaluations, Shown Under Different Conditions of Subordinate Gender, Evaluator Gender, and Evaluator SDO (Study 2).

Discussion

Study 2 replicated the basic pattern that emerged in Study 1 but only for male evaluators whose SDO was high. This suggests that although the predicted effect received marginal support, each of the predicted evaluator moderators alone is not enough to elicit the effect. Importantly, these evaluators are most similar to the naturally occurring military sample we looked at in Study 1: Military personnel tend to be higher on SDO (Nicol et al., 2007), and 97% of the commanding officers in Study 1 were male.

In addition, because the effect only emerged among those participants who theoretically would be most threatened by a strong competence-signal female, this finding is consistent with our proposed mechanism: threat to the gender hierarchy. As predicted, none of these variables (evaluator gender, evaluator SDO, or subordinate competence signals) affected evaluations of male subordinates. Thus, it does not appear to be a threat to a general hierarchy that is being reacted to; rather it is a threat to the gender hierarchy specifically.

At the same time, the four-way interaction and the specific contrast of interest (the effect of competence signal strength on performance evaluations for female subordinates being evaluated by male, high SDO supervisors) were only marginally significant. Although a conceptually similar contrast was significant in Study 1, these interpretations should be treated with caution. In the next study, we conceptually replicate the design to further test the robustness of the effect.[6]

Studies 3a and 3b

The primary purpose of Studies 3a and 3b was to demonstrate that the type of questions used in a performance evaluation can help reduce the bias that emerged in Studies 1 and 2. Following the logic outlined in the “Reducing Evaluator Bias” section, we attempted to replicate the effects that emerged in Studies 1 and 2 with a subjective measure of performance but show that it is minimized or eliminated when a more objective measure is used.

Two additional goals of Studies 3a and 3b were to further test the boundary conditions of the results that emerged in Study 1 and provide evidence for mechanism. In Study 2, we tested the moderating role of subordinate gender, evaluator SDO, and evaluator gender. Here, we wanted to test the effect of subordinate on-the-job objective performance (Hypothesis 3). To avoid creating an overly complex design with a five-way interaction (see Halford, Baker, McCredden, & Bain, **2005**), and because the effect was most pronounced in Study 2 for male evaluators who were high on SDO, we selected male evaluators for Study 3a and female evaluators for Study 3b. The resulting design for each study was a 2 (subordinate competence signal) × 2 (subordinate gender) × 2 (subordinate on-the-job objective performance), with evaluator SDO included as a continuous variable. We manipulated the quality of the subordinate's on-the-job performance (i.e., the work contributions being evaluated) as either low or high by varying the number of corrections made to a document containing many mistakes.

A final goal of Studies 3a and 3b was to replicate the pattern of results that emerged in the prior studies but with a different competence signal manipulation. Revealing the predicted effect with different competence signals provides evidence for the robustness and generalizability of the phenomenon. Thus, we used prior job performance as a competence signal in Studies 3a and 3b. All participants (i.e. evaluators) in Studies 3a and 3b were told that they were salespeople in a software firm, that their booked sales for the prior year was £84K, and that their subordinate booked either a similar (high competence signal) or a lower level (low competence signal) of sales in a prior job in an unrelated industry.

Study 3a

Method

Participants

One hundred twenty-two male adults participated in this laboratory experiment conducted at a business school in London, United Kingdom. Participants’ mean age was 29.07 years (*SD* = 9.99). They were 39.3% of European ethnicity, 29.5% of Asian ethnicity, 16.4% of African ethnicity, .8% of

Native American ethnicity, and 13.9% selected “Other” for their ethnic background. The majority (79.5%) had at least an undergraduate degree. Seventy-one percent listed English as their first language. Twenty percent reported being employed full-time, and 69% reported being a student. They were recruited via a discussion list maintained by the business school that is made up of individuals willing to come into the laboratory and be paid to complete studies. They were paid £10 for their time.

Design and procedure

The experiment was a 2 (subordinate competence signal: weak [low former sales] vs. strong [high former sales]) × 2 (subordinate gender: male vs. female) × 2 (subordinate on-the-job objective performance: low [few corrections] vs. high [many corrections]) with participants’ SDO as a continuous factor.

Upon arrival in the lab, participants were told that they would be acting as a senior sales manager in a software development firm called Solv|Sys plc. They were also told that they had booked £84K in sales last year and managed two sales associates. Next, they were asked to turn over a piece of paper on their desk, which was an organizational chart that showed them as a manager with two subordinates—they were asked to circle their own role.

Then, they were told that they would be working with one of the associates—either Michael (*male subordinate*) or Claire (*female subordinate*), depending on condition—and were asked to circle this person on the organizational chart. The subordinate's position was clearly lower on the hierarchy than the participant's own. Participants were then given background information on the subordinate, which contained the competence signal manipulation. The subordinate was said to have joined Solv|Sys earlier this year. He or she had previously worked for a company that makes custom-made wooden furniture and had booked either £28K in sales last year (*weak competence signal*) or £78K in sales last year (*strong competence signal*).

Participants were then told that they had to submit an important sales pitch for another senior sales manager who was supposed to submit the pitch while traveling abroad, but his computer had been stolen. The only record of the pitch was an early hard-copy draft that had been left on his desk. The participant was told that he had been in a meeting for the past hour and had asked the subordinate to start editing the rough draft. At this point, the participant was prompted to ask the experimenter for the hard copy, which contained the “subordinate's” corrections in pen. We manipulated the hand-written corrections to either be in stereotypically male or female script (consistent with the subordinate gender manipulation), and we manipulated the number of corrections to be either low or high, consistent with condition. In the *low on-the job performance* condition, 12 edits were made to the document. In the *high on-the-job performance* condition, 40 edits were made.

Upon receiving the hard-copy of the pitch, participants were told that they would have 10 minutes to type the pitch into the prompt on the computer, including whatever edits they saw fit. The pitch would automatically be submitted after 10 minutes. After submission, participants evaluated their subordinate's performance.

Dependent measures

Participants evaluated their subordinate in two ways. First, we asked them to evaluate the subordinate on more subjective dimensions by answering the following questions: “Please rate Michael's [Claire's] performance using the scale below” (seven-point scale: 1 = *extremely bad* , 7 = *extremely good*), “How proactive would you say Michael [Claire] is?” (five-point scale: 1 = *not at all* , five = *extremely*) and “How much initiative would you say Michael [Claire] displayed?” (five-point scale: 1 = *very little* , 5 = *a great deal*). These questions were combined to form a subjective evaluation composite (α = .86, first item transformed to five-point scale).

Second, we asked participants to evaluate Michael [Claire] on an objective dimension by having them answer the following question: “How many errors in the draft pitch did Michael [Claire] catch?” (1 = *almost none of them* , 3 = *about half of them* , 5 = *almost all of them*).

Last, we asked participants to respond to the SDO scale (described in Study 2), provide demographic information, and answer two attention check questions (“What was your subordinate's gender?” and “Your booked sales last year were:” with the response options: “ *Far less than your subordinate's* ,” “ *About the same as your subordinate's* ,” “ *Far greater than your subordinate's* ,” and “ *I don't know* ”).

Results

Subjective evaluation

We tested our predictions using hierarchical multiple regression (Cohen et al., **2003**). The overall effects of subordinate competence signal (0 = *weak* , 1 = *strong*), subordinate gender (0 = *male* , 1 = *female*), subordinate on-the-job performance (0 = *low* , 1 = *high*), and evaluator SDO (mean-centered) were entered into the first block, the two-way interaction terms were added to the second block, the three-way interaction terms were added in the third block and the four-way interaction term in the last block.

Step 1 revealed a significant effect of subordinate gender, (β = .17), *t* (117) = 2.13, *p* = .036, and of on-the-job performance, (β = .42), *t* (117) = 5.16, *p* < .001. Steps 2 and 3 revealed no significant interactions. In Step 4, the four-way interaction was significant (β = –.48), *t* (106) = –1.99, *p* = .050. See Table 4 for step-wise regression results.

Table 4. Regression Results From Study 3a

	Step 1			Step 2	
Variable	**b**		**95% CI**	**b**	
Subordinate gender[a]	0.32[*]		[0.02, 0.62]	0.59[*]	
Subordinate past accomplishment (sales figure in past job)[b]	−0.15		[−0.45, 0.15]	0.15	
Subordinate on-the-job performance[c]	0.78[**]		[0.48, 0.15]	1.05[***]	
Evaluator SDO[d]	−0.08		[−0.23,	0.14	

		-0.08]		
ΔR^2	0.23			
F (ΔR^2)	8.47***			
Subordinate gender X Subordinate past accomplishment			−0.31	
Subordinate gender X Subordinate on-the-job performance			−0.18	
Subordinate gender X Evaluator SDO			−0.18	
Subordinate past accomplishment X Subordinate on-the-job performance			−0.34	
Subordinate past accomplishment X Evaluator SDO			0.00	
Subordinate on-the-job performance X Evaluator SDO			−0.20	
ΔR^2				0.04
F (ΔR^2)				0.99
Subordinate gender X Subordinate past accomplishment X Subordinate on-the-job performance				
Subordinate gender X Subordinate past accomplishment X Evaluator SDO				
Subordinate gender X Subordinate on-the-job performance X Evaluator SDO				
Subordinate past accomplishment X Subordinate on-the-job performance X Evaluator SDO				
ΔR^2				
F (ΔR^2)				
Subordinate gender X Subordinate past accomplishment X Subordinate on-the-job performance X Evaluator SDO				
ΔR^2				
F (ΔR^2)				
R^2	0.23***			0.26**

Note

Stepwise regression model. *F* -test indicates model fit increase. CI = confidence interval.

a 0 = male, 1 = female.

b 0 = much lower than participant, 1 = almost equal to participant.

c 0 = male, 1 = female.

d Mean-centered.

+ $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

We interpreted the 4-way interaction the same way as in Study 2. The only set of conditions that showed a significant negative relationship between competence signal strength and performance evaluations was high-performing female subordinate being evaluated by high SDO male evaluators, ($\beta = –.45$), $t(106) = –2.19$, $p = .031$. All other conditions yielded *p* s > .15, indicating that the relationship between competence signal strength and performance evaluations for male subordinates was not significant under any condition. See Figure 3 for a depiction of the weak versus strong competence signal slope across each of the eight experimental conditions. Neither current on-the-job performance, $F(1, 60) = 1.41$, $p = .240$, $\eta_p^2 = .023$, nor evaluator SDO, ($\beta = –.00$), $t(60) = –.004$, $p = .997$, alone was enough to elicit the negative relationship between competence signal strength and performance evaluation for female subordinates.

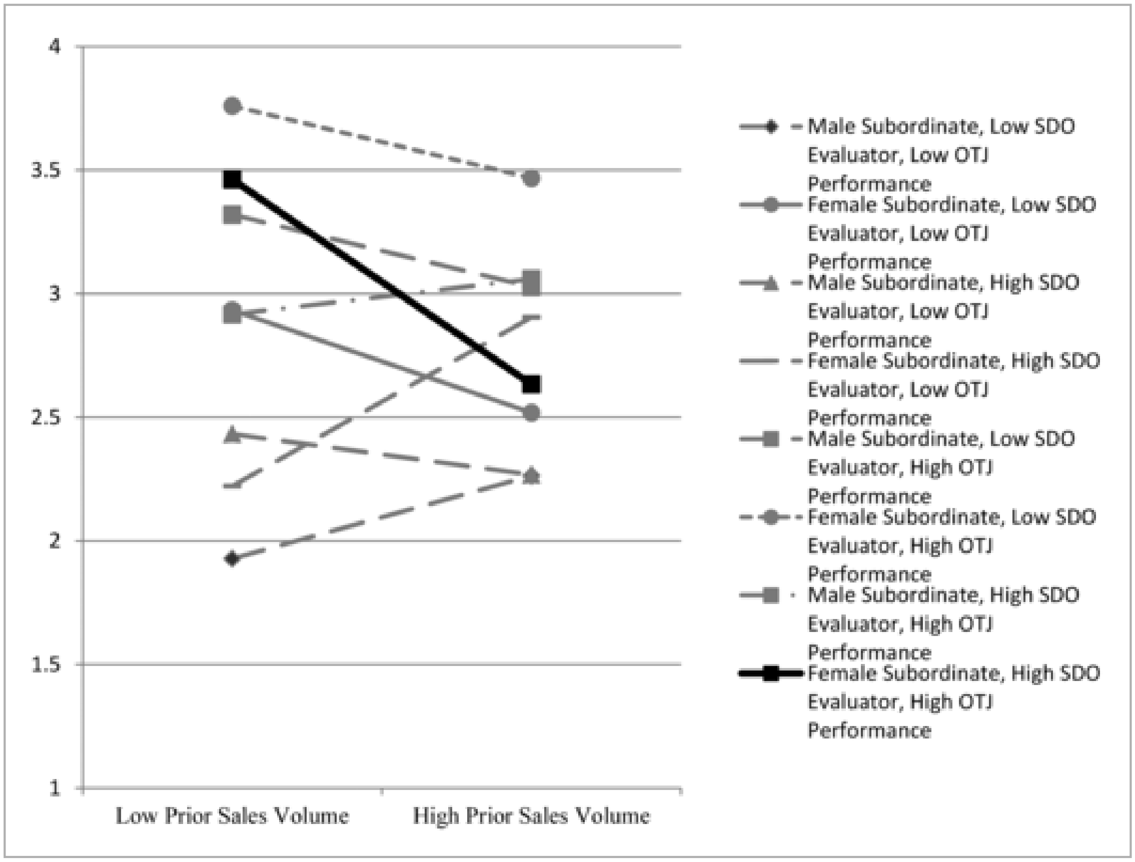

Figure 3.

Open in figure viewer

The Effect of Subordinate Competence Signal (Sales Volume in Prior Job) on Performance Evaluations by Male Evaluators, Shown Under Different Conditions of Subordinate Gender, Subordinate On-the-Job (OTJ) Performance and Evaluator SDO (Study 3a).

Objective evaluation

We next tested our predictions using the same model as described above but with the objective evaluation (i.e., number of errors corrected) as the dependent measure. Step 1 revealed a significant effect of subordinate on-the-job objective performance, (β = .64), *t* (117) = 9.00, *p* < .001. Steps 2, 3, and 4 revealed no significant effects. As expected, those subordinates who did an objectively better job in finding errors were given a more positive evaluation, and this relationship was not moderated by any other variable.

Study 3b

Method

Participants

One hundred forty-three female adults participated in this laboratory experiment conducted at the same business school in London, United Kingdom. Participants' mean age was 27.32 years (*SD* = 8.06). They were 37.8% of European ethnicity, 24.5% of Asian ethnicity, 15.4% of African ethnicity, 1% of Native American ethnicity, and 21.7% selected "Other" for their ethnic background. The majority (83.1%) had at least an undergraduate degree. Eighty-five percent listed English as their first language. Thirty-three percent reported being employed full-time, and 50% reported being a student. They were recruited via a discussion list maintained by the business school that is made up of individuals willing to come into the laboratory and be paid to complete studies. They were paid £10 for their time.

Design, procedure, and dependent measures

These were identical to Study 3a, except that all participants were female.

Results

Subjective evaluation

We tested our predictions using the same model as described in Study 3a. Step 1 revealed a significant effect of subordinate gender, (β = –.16), *t* (138) = –2.06, *p* = .042, and of on-the-job performance, (β = .43), *t* (138) = 5.63, *p* < .001. In Step 2, the current on-the-job performance × competence signal interaction approached traditional levels of significance (β = –.27), *t* (132) = –1.95,

p = .054. The pattern suggested that for those who performed poorly on the job, stronger competence signals had no effect on evaluations, t < 1, whereas for those who performed well on the job, stronger competence signals tended to lead to more negative evaluations, (β = –.17), t(132) = –1.21, p = .228, although this effect was not significant.

Objective evaluation

As in Study 3a, we tested our predictions for the objective performance evaluation item using the same model as described above but with the objective evaluation (i.e., number of errors corrected) as the dependent measure. Step 1 revealed a significant effect of subordinate on-the-job objective performance, (β = .56), t(137) = 7.88, p < .001. No significant effects emerged in Step 2. Step 3 revealed a significant effect for the SDO × subordinate gender × current on-the-job performance, (β = –.58), t(127) = –2.39, p = .018, such that for low SDO evaluators, the subordinate gender × current on-the-job performance interaction was null, t < 1, whereas for high SDO evaluators it was significant, (β = –.56), t(127) = –2.32, p = .022. In this interaction, male subordinates benefited from stronger current on-the-job performance, (β = .91), t(127) = 4.28, p < .001, whereas the same was not true for female subordinates, (β = .31), t(127) = 1.70, p = .091. Although not part of our theorizing or predictions, it is interesting to note that women who tend to support the maintenance of hierarchies reward men for greater on-the-job performance more than they do women. Such actions are supportive of gender hierarchy maintenance, even though it is to the disadvantage of the evaluators' own gender. This is consistent with existing work, which demonstrates that both men and women can hold sexist beliefs and show bias against women.

There was also a marginal effect for the subordinate gender × current on-the-job performance × competence signal strength interaction, (β = –.35), t(127) = –1.75, p = .082. The pattern indicated that although the subordinate gender × competence signal strength effect was null for those who had performed poorly on the job, t < 1, this same interaction was stronger for the high on-the-job performer, (β = .40), t(127) = 2.36, p = .020. High-performing male subordinates received worse evaluations the stronger their competence signal, (β = –.35), t(127) = –2.38, p = .019, whereas female subordinates were not affected by competence signal strength, t < 1. At the risk of overinterpreting a finding that emerges from a marginal three-way interaction, it is interesting to note that in this case high-performing male subordinates were disadvantaged as a result of stronger competence signals, whereas high-performing female subordinates were not. Although it is difficult to account for this result without mechanism information, it is possible that female evaluators felt threatened by high-performing, high past competence signal male subordinates. If true, then it is surprising that the effect emerged on more objective evaluations but not on more subjective evaluations.

The four-way interaction was not significant, (β = –.31), t(126) = –1.06, p = .294.

Discussion for Studies 3a and 3b

These findings replicate Studies 1 and 2, and show that for a woman who is objectively doing a good job at work and has a male, high SDO evaluator, the stronger her competence signal the worse performance evaluations she will receive. We also found that this particular negative bias was eliminated with a more objective measure.

General Discussion

Advanced degrees, high ranking in a graduating class, a history of fast-track promotions—hopeful job applicants document these accomplishments on their curriculum vitae with the expectation that they will help them obtain greater future career success (e.g., Becker, 1964 ; Judge, Cable, Boudreau, & Bretz, 1995). Not only can these achievements positively affect employers' hiring decisions, but theory suggests that such competence signals also may impact performance evaluations down the line by positively biasing evaluators' broader beliefs about an employee (Berger et al., 1972 ; Cialdini, 1993 ; Rosenthal & Jacobson, 1966).

Unfortunately, the story can have a different ending for female employees because competence signals may come back to haunt them in the future. By integrating signaling theory with the concept of gender hierarchy threat, we hypothesized that evaluators would give worse performance evaluations to women with stronger competence signals compared to women with weaker competence signals because the former are status incongruent. We found evidence consistent with this prediction using three different operationalizations of competence signals. Further, in Studies 2, 3a, and 3b, the competence signals were completely unrelated to the work being evaluated (which was held constant), and still this negative bias emerged. Results strongly suggested that the reason why women received biased evaluations for stronger competence signals is that evaluators are threatened by their status incongruence.

Contributions: Applied and Theoretical

Perhaps the most important contribution of this paper is exposing an important third hurdle faced by women in the workplace. Research has shown that women are perceived as less competent (the first hurdle), and to overcome this they may try to use male-stereotyped behaviors (e.g., self-promotion). Research also has shown this leads to a second hurdle: These male-stereotyped behaviors are normatively prohibited for women (Heilman & Haynes, 2005 ; Heilman, Wallen, Fuchs, & Tamkins, 2004 ; Rudman & Glick, 1999), leading to backlash. Another way that women can try to overcome the perception of reduced competence is through visible accomplishments, which should offer honest signals of greater competence. However, our research sheds light on competence signals as a third hurdle: The very accomplishments that help a woman get hired may lead to lower performance evaluations down the line. Specifically, our research shows that some supervisors may perceive stronger competence signals as a threat to the gender hierarchy and provide lower performance evaluations as a result.

By focusing on performance evaluations of a subordinate by his or her supervisor, our results also broaden the existing understanding of how gender-norm deviant behavior affects women's career outcomes because existing research has focused on women's transitions into leadership roles (Eagly & Karau, 2002 ; Heilman & Parks-Stamm, 2007 ; Okimoto & Brescoll, 2010 ; Rudman & Glick, 2001 ; Rudman et al., 2012). Although clearly it has been valuable to expose gender discrimination at leadership transition points, it also is likely that transitions into leadership roles are particularly vulnerable to punishment. This is because leadership positions pose a special threat to the traditional gender hierarchy (Rudman et al., 2012), and transitions into leadership represent mobility striving (Rosenbaum, 1979), which exacerbates the threat. As such, our research provides a more conservative test of the general theory that norm-deviant women are penalized at work. Moreover, by revealing this negative bias in performance evaluations, we show how generalized the backlash effects for competence signals are. After all, performance evaluations are a very routine organizational process that affects most women's pay and advancement—not just those striving for leadership positions.

Our findings provide a theoretical contribution to signaling theory because they suggest an unexpected negative externality of competence signals. An assumption of signaling theory is that individuals on the job market benefit from signaling competence through achievements that would be harder for less qualified people to accomplish. Although the direct effect of honest signals on hiring success has received copious support, the indirect effect of knowledge about competence signals on future career outcomes is less well understood. Past research has not exposed the possibility that the very signals that are helpful during the hiring process might be harmful in the evaluation process, at least for individuals who threaten the status hierarchy.

Our results also represent a theoretical contribution to the status incongruence hypothesis, in that we show how reactions to status incongruence operate in much more subtle, but possibly more damaging, ways than previously theorized. Existing work in this area has predominantly focused on intergender comparisons, demonstrating that agentic men receive better outcomes than agentic women because the latter receive a dominance penalty. The within-women effect, which compares career outcomes for agentic versus nonagentic women, has received less attention, presumably because it is counterintuitive that a more competent woman would receive worse career outcomes than a less competent woman. Here we demonstrate that this exact pattern can emerge when the competence signal is long in the past. For example, results in Study 2 demonstrated that, for a female subordinate, certain evaluators evaluated the same output as worse when she had a college education versus a high-school education. This is particularly noteworthy because the participant was simply asked to evaluate a list of ideas generated by the subordinate, and we did not highlight the subordinate's degree as being relevant to the task nor did we suggest that the educational degree was particularly high or low. Rather, it was included in what was described as background information on the subordinate.

Study 2 also served to rule out norm violation as the mechanism driving this process. If the mechanism were norm violation for competent women, then differences between evaluators such as SDO and gender should theoretically not moderate the effect. Because these variables do moderate the effect in the predicted ways, this suggests that gender hierarchy threat is the key mediator. In this way, our study builds on the status incongruence hypothesis (Rudman et al., 2012) to predict that—because threat drives this process—certain features of evaluators should make them particularly threatened by women's accomplishments and thus more likely to evince backlash.

Finally, our results are interesting to consider relative to research on expectancy violation theory (Jussim, Coleman, & Lerch, 1987) and the shifting standards model (Biernat & Kobrynowicz, 1997), which predict that women can benefit more than men for their achievements. Although these findings are mostly concerned with comparing evaluations of male versus female targets at a given level of achievement (although our predicted effect is a within-female one), it is nevertheless interesting to consider how the previous findings might be related to ours. For example, in Study 1 we found that a woman whose competence signal was proximal to her evaluator's received lower evaluations compared to a man at the same level of proximity to his evaluator, which is conceptually at odds with expectancy violation and shifting standards. Indeed, it is more in line with status characteristics theory, which predicts that women need to demonstrate greater evidence of achievement compared to men in order to be judged as competent (Berger et al., 1972 ; Foschi, 1992). One reason that our pattern of effects is more consistent with status characteristics theory may lie in the type of evaluation. Biernat and Kobrynowicz (1997) demonstrated that women tend to receive higher ratings than men on judgments of minimum competency whereas they receive lower ratings than men on judgments of general ability, such as test grades. Because the evaluators in our studies are essentially rating the general ability level of the targets, then women should be disadvantaged relative to men.

Future research might investigate when these different responses to women's accomplishments may arise and how they might interact. For example, it is possible that an evaluator might see a female subordinate's strong past competence signal as an expectancy violation and judge her past accomplishment as noteworthy, which then evokes gender hierarchy threat and leads to a downgrading of current performance. In addition, it may be worth considering how evaluating a work subordinate, rather than an anonymous target, changes the manner in which judgments are made. Past research in expectancy violation and shifting standards tends to rely on judges that are not dependent on the target. In contrast, the evaluators in our studies required the input of their subordinates and were thus dependent on them. Does dependence on the target of evaluation change the way judgments are made in predictable ways? For example, it is possible that greater dependence would make evaluators more likely to compare a subordinate's performance against a universal standard rather than against a gender-specific one (see Biernat & Kobrynowicz, 1997) because they are mainly concerned with how much they benefited egocentrically from the subordinate. Future research could test this prediction.

Practical Implications

Our results have several practical implications, because they shine light on a hurdle that many employees must overcome, and they broaden our understanding of how evidence of success affects women's career outcomes. Because these discrimination effects are counterintuitive, it is only by making them public that employers can understand how women are evaluated when the process is left unchecked. First, these findings have critical implications for women who aspire to achieve professional success. Many women feel overwhelmed by—or simply don't like—the idea that they should act more like men to succeed in their careers. As a result, they turn to concrete accomplishments as a way of circumventing such pressures. However, these results show that this alternative also carries negative consequences down the line. In fact, our results suggest that this penalty may increase as women become more senior in their organizations and their track record of accomplishments becomes increasingly evident and increasingly threatening. Indeed, this may be an additional dynamic that prevents women from breaking through the glass ceiling of high-powered positions in organizations. Of course, at an individual level we do not believe that these findings should discourage women from accomplishing in the first place, or from highlighting their past accomplishments. However, at a societal level, we hope that by uncovering and broadcasting this dynamic we can help leaders become more aware of these trends and make them less likely to hire those who would be biased.

The results of our studies have other important implications for organizations. Given that firms' competitive advantage springs from both workforce human capital (e.g., advanced degrees from reputable universities) and strong on-the-job performance (Hitt, Bierman, Shimizu, & Kochhar, 2001), it may be possible and competitively advantageous for employers to try and reduce the role that gender and backlash effects play in their assessment decisions. At the same time, our results suggest that not all evaluators are prone to this negative bias. Only male evaluators who are high on SDO and evaluating a high-performing female subordinate appear to be prone. This implies that some industries and departments (e.g., the police) may be more vulnerable to this effect than others (e.g., public defenders) because the evaluators are more likely to display these characteristics (Sidanius, Liu, Shaw, & Pratto, 1994). For such groups, a variety of avenues exist that would help to reduce the negative bias.

First, organizations should strive to include objective measures (which focus on "direct measures of countable behaviors or outcomes," Bommer, Johnson, Rich, Podsakoff, & Mackenzie, 1995 , p. 588) of employees' work output and impact when evaluating their performance. Despite the prevalence of subjective performance evaluations in organizations (which depend on "supervisor ratings of employee performance"), it is generally acknowledged that they are unsound because they are not linked to company strategy, and they often do not help employees understand how to improve (e.g., Bommer et al., 1995 ; Schmidt & Hunter, 1992). Further, research shows that forming accurate performance evaluations is difficult and subject to a number of cognitive biases (DeNisi & Williams, 1988). We found in Study 3a that the negative bias associated with gender hierarchy threat disappeared with a more objective measure of performance. Thus, our research adds to the list of reasons for organizations to strive toward objective measures of employees' work output and impact when evaluating their performance.

It is not feasible to completely eradicate subjective performance measures from evaluations, however. For example, most work is interdependent to some extent, and it is difficult to isolate a given employee's personal contributions (e.g., creativity, insights), as well as the value of those contributions to the organization (Landy & Farr, 1980). However, there are ways to reduce bias even when using such measures. For example, evidence suggests that educating people about stereotypes and prejudices can reduce implicit stereotypes, which occur on an automatic, unconscious level (Rudman, Ashmore, & Gary, 2001). As suggested by Duehr and Bono (2006), if such associations can be unlearned through training, future research may confirm that training helps eliminate the backlash effects revealed in these studies. For example, it may be possible for organizations to help leaders increase awareness about their own unconscious assumptions using an implicit stereotype program (Greenwald & Banaji, 1995) where participants sort words into categories that take more cognitive processing and more time when a word (e.g., woman vs. man) violates a stereotype (e.g., competent).

Another approach that organizations may consider implementing in light of our results is to separate hiring from evaluation functions. Although some organizations rely heavily on human resource departments to drive hiring processes, many organizations encourage hiring to be performed by the same supervisors who will manage and evaluate the hired individuals. This approach certainly has benefits because those doing the hiring are particularly well informed about the nature of the work. However, the findings presented here also suggest a potential limitation of this strategy because it appears that knowledge of applicants' past accomplishments, which should have no bearing on future performance evaluations, may nevertheless lead to negative biases for female employees. Therefore, our results offer a new lens on hiring that suggests gathering input from future supervisors through one-on-one interviews (e.g., case interviews for consulting jobs) but shielding them to the extent possible from the knowledge of past career accomplishments and other competence signals.

Limitations, Strengths, and Future Directions

Despite the contributions of this research, there are limitations that should be noted. First, our research did not directly measure perceived threat nor did we examine the extent to which the resulting punishment was conscious or unconscious. This is because supervisors rarely admit feeling threatened by women with greater status. However, we did show moderation by three variables that theoretically should exacerbate the level of gender hierarchy threat experienced by an evaluator—evaluator gender, evaluator SDO, and subordinate on-the-job performance—which is consistent with our proposed mechanism. Indeed, this method of demonstrating process has been proposed as superior to direct measurement of mediation model when the process variable is difficult to measure (Spencer et al., **2005**), as is the case in this paradigm. Nonetheless, it would be interesting for future research to test threat more directly (e.g., brain scan activity, physiological measurement) or use qualitative research to examine whether supervisors articulate evidence of their motivations. Moreover, future research may reveal that in order to be threatening, a woman's competence signals need to be in a male-typed domain such as business success (e.g., it is less likely that even high social dominance men would feel threatened by a woman's past achievements in nursing).

Next, our field investigation focused on a single organization in which gender differences are made salient by the fact that women have historically been prohibited from holding certain key roles. As a result, evaluators in this environment would presumably maintain a strong gender distinction. Even though the U.S. Armed Forces is one of the world's largest employers, and even though many nonmilitary women work for male supervisors who support the existing status hierarchy, it nevertheless would have been useful to examine multiple organizations and confirm the moderating effect of gendered environments. Although the participants in our online and laboratory-based samples were heterogeneous and likely held many different types of jobs, we did not collect data on their industry, jobs, or income level, and thus we could not control for these variables. It will be interesting to examine whether the results presented here generalize better to some industries than others due to how gendered they are (e.g., manufacturing vs. consulting). Likewise, it would be interesting to examine these same hypotheses in groups and organizations where women are the majority. As noted by Vecchio and Brazil (**2007**), when women comprise a smaller portion of a work group they are more likely to receive lower evaluations; but when they constitute a majority, their evaluations may exceed those obtained by men.

Although we operationalized competence signals in three ways (pay level, education level, and performance level on unrelated jobs), it would have been useful to investigate whether the same effect emerged for other competence signals. For example, similar effects might emerge from the prestige of one's degree-granting institution or from the reputation of one's prior employers. Thus, another area for future research would be to investigate the types of competence signals that yield the demonstrated effect. In addition, this investigation focused on the gender hierarchy and discrimination against women. Our theory suggests, however, that a similar bias might emerge for other groups that are stereotyped as lower status and less competent, such as certain ethnic minorities or lower socioeconomic groups (e.g., see Fitzsimons & Lehmann, 2004 ; Word et al., 1974). Although none of our samples had enough individuals from ethnic minority groups to test this possibility, it represents a promising avenue for future research.

In our studies subordinates' competence signal never outstripped that of their evaluator—rather they were either far weaker or approximately equal. Thus, we do not know how evaluators would respond to subordinates with stronger competence signals than their own. Future research may reveal that these evaluators would be even more threatened and provide even more negative performance evaluations. On the other hand, the effect may plateau, such that any subordinate competence signal that approaches the evaluator's own is threatening but the threat does not linearly increase beyond a certain point.

In addition, although we found results consistent with our predictions in the domain of performance evaluations, we did not compare this effect against other career outcomes such as hiring, raises, or bonuses. Conceptually, the negative bias we document should be more likely to emerge the more distal the competence signal is to the career outcome. Past competence signals are very proximal to hiring decisions because, prescriptively, people with stronger evidence of past successes are those that are more likely to be hired. Thus, even though recruiters may experience gender hierarchy threat, they may be less likely to evince a negative relationship between competence signals and hiring success. In contrast, past competence signals are not proximal to performance evaluations, in that there is no prescriptive relationship between them. This allows the negative bias to emerge. It seems possible that a similar negative relationship would emerge for other career outcomes for which competence signals are distal, such as raises and bonuses.

Finally, we did not systematically account for the comparison group when asking participants to evaluate their subordinate's performance. Prior research (e.g., Biernat & Kobrynowicz, 1997) has shown that the type of scale used in evaluations can affect whether participants generate a within-group comparison (e.g., "this woman is very competent, compared to most women") or a cross-group comparison (e.g., "this is a competent person"). The within-group comparison can yield surprising counterstereotypical results (Jussim et al., 1987), with women receiving higher scores than men on male-type tasks. Our studies are different from past research, in that evaluators were predominantly responding from a self-centered perspective (e.g., "this person was critical to my decision making" and "this person contributed to the final advice submitted"), and were dependent on the subordinate's input. It seems likely that our approach would be more likely to generate cross-group rather than within-group comparisons because the relevant metric is how much the work helped the respondent. It would be interesting for future research to systematically test the effects of evaluation wording (open or closed to within-gender comparison) and evaluator-target interdependence.

Our investigation also has some clear strengths. First, our field investigation established the existence of our phenomenon in actual employment situations. Much of the existing "backlash" research has been conducted in laboratory experiments with undergraduate students who have no structural relationship with the women being evaluated. Showing "backlash effects" on real employees is an important contribution because students can show different beliefs about women than managers (Duehr & Bono, 2006). Moreover, to help rule out alternative explanations for the results such as priming or mood effects, we gathered different types of field data from multiple sources (e.g., we combined archival records of employees' sex and pay grades with supervisors' performance evaluations). As such, Study 1 avoided many of the common confounds with field research (e.g., common method variance, priming, hypothesis guessing) while maximizing external validity.

Then, we paired these field study results with two experiments in order to permit strong implications about causality. The experiments also allowed us to illuminate the conceptual mechanism at work, as well as control for actual performance in order to make strong inferences about the observed negative bias. In sum, we replicated our results across samples, methods, and types of competence signals because there is "tremendous value in conducting multiple tests of the same phenomenon" (Tepper et al., 2009 , p. 165), in that it provides evidence of robustness.

Conclusion

Many women focus on obtaining career-relevant competence signals because they assume it will lead to greater career success. In this research, we explore the possibility that although stronger competence signals may help them land a job, they may also come back to haunt them in the future, leading to lower performance evaluations. By nature, competence signals build across an employee's career, meaning that positive performance ratings can become harder and harder for women to obtain as their careers progress. Thus, our paper helps shed new light on the glass ceiling that tends to keep women out of top management positions.

[1] We did not have age data for 35 legal advisors, so age is not included in the correlation matrix.

[2] The pattern of results is the same when female supervisors are included in the dataset.

[3] Some supervisors evaluated the same subordinate. We initially ran a hierarchical linear model to account for this characteristic, with subordinate as the higher-level identifier. When we compared this model to the single-level OLS regression, there was no significant improvement and thus we have reported the simpler regression results.

[4] Regressing subordinate evaluation onto subordinate pay grade, subordinate gender and their interaction, and controlling for subordinate tenure in military, yielded a null interaction, $t<1$, suggesting that evaluator pay grade needs to be taken into account for the predicted effect to emerge.

[5] A 2 (subordinate gender) × 2 (subordinate past accomplishment) × 2 (evaluator gender) ANOVA revealed that none of these variables nor their interactions affected SDO scores, all p s > .1.

[6] Given the marginal results for the key contrast of interest in this study, we later test for the robustness of the effect by performing a meta-analysis on the effect of low versus high

competence signals for high-performing female subordinates when evaluated by male, high SDO evaluators across all three studies. From Study 1, we focused on the low versus high proximity contrast for female subordinates (the legal advisors are a high-performing group, all the evaluators were male, and because they are in the military, likely higher than average on SDO). From Study 2, we focused on the low versus high education contrast for the female subordinate being evaluated by male, high SDO evaluators. From Study 3a, we focused on the low versus high past sales performance contrast for the female subordinate who was a high on-the-job performer and being evaluated by male, high SDO evaluators. We used the methods described by Rosenthal and Rosnow (**1991**) to calculate the combined *p* -level and the combined effect size across these studies. The combined *p* -level for this specific contrast is $p < .0003$ (unweighted) or $p < .0008$ (weighted by *df*), and the combined effect size is $r = .18$ (unweighted) or $r = .17$ (weighted by *df*). This meta-analytic result suggests that the effect sizes are robust and significant.

REFERENCES

Wiley Online Library

Help **Browse by Subject** **Browse Publications** **Resources**

Agents | Advertisers | Cookies | Contact Us | About Us

Privacy | Site Map | Terms & Conditions | Media

WILEY

Wiley.com **About Wiley** **Wiley Job Network**