


Shifting Standards and the Inference of Incompetence: Effects of Formal and Informal Evaluation Tools

Personality and Social
Psychology Bulletin
XX(X) 1–14
© 2010 by the Society for Personality and
Social Psychology, Inc
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146167210369483
http://pspb.sagepub.com


Monica Biernat¹, Kathleen Fuegen², and Diane Kobrynowicz³

Abstract

The authors distinguish between minimum and confirmatory standards of incompetence and hypothesize that for groups stereotyped as relatively competent (or deficient in incompetence), minimum standards of incompetence are lower (suspicion of incompetence is triggered sooner) but confirmatory standards are higher, relative to groups stereotyped as relatively incompetent. An initial study demonstrated this evidentiary pattern for male versus female targets. In Studies 2 and 3, participants were exposed to a poor-performing male or female (Study 2) or Black or White male (Study 3) trainee and were asked to record “notable” behaviors in either their “informal notes” (instantiating a minimum standard) or a “formal performance log” (instantiating a confirmatory standard). Consistent with predictions, fewer incompetent behaviors were recorded in the formal log than in informal notes for White male trainees. Firing decisions generally mimicked these patterns and in Study 3 were partially mediated by the accessibility of incompetent behaviors.

Keywords

stereotypes, social judgment, competence, bias

Received November 5, 2008; revision accepted November 19, 2009

One message that emerges from the research literature on stereotyping and prejudice is that negative outcomes for the stigmatized often may take subtle rather than blatant forms. For example, research on “aversive racism” suggests that negative perception and treatment of Blacks may occur only in situations where nonracial reasons for the mistreatment are available and that bias can take such forms as the differential weighting of information about Black versus White targets (Dovidio & Gaertner, 2000; Gaertner & Dovidio, 1986; Hodson, Dovidio, & Gaertner, 2002). With regard to gender, the form of bias may not be across-the-board discrimination against women in the workplace but rather backlash against women who are too agentic, or too successful as managers (Heilman, 2001; Rudman, 1998; Rudman & Fairchild, 2004).

In research from our own lab, we have also argued that the specific *pattern* of stereotyping of the stigmatized is complex and may depend on a number of factors. First, judgments of stereotyped group members may show assimilation effects (targets judged consistently with stereotypes) or contrast effects (targets judged in a counterstereotypical direction), depending on the form of judgment. In general, judgments of others made on *common rule* scales (e.g., rank orderings, estimates of monetary worth, or standardized test scores)

reveal assimilation to stereotypes, whereas judgments made on *subjective* response scales (e.g., semantic differentials, trait ratings) may reveal reductions or reversals of these effects (for reviews, see Biernat, 2003; Biernat & Manis, 2007).

Second, members of stereotyped groups may be held to either low or high standards depending on the *type* of standard being assessed. *Minimum* standards are expectations for a group and tend to directly reflect stereotypes; thus, minimum standards tend to be *lower* for group members stereotyped as deficient on a given attribute. *Confirmatory* standards, on the other hand, are thresholds that reflect *certainty* that an individual has an attribute; these standards tend to be *higher* for group members stereotyped as deficient. For example, minimum standards for competence in “masculine” occupations are *lower* for women than men, but confirmatory standards

¹University of Kansas, Lawrence, KS, USA

²Northern Kentucky University, Highland Heights, KY, USA

³University of Texas at Austin, Austin, TX, USA

Corresponding Author:

Monica Biernat, University of Kansas, 1415 Jayhawk Blvd., 426 Fraser Hall,
Lawrence, KS 66045-7556
Email: biernat@ku.edu

(how much evidence would it take to be certain that a woman is competent?) are *higher* for women than men, and for Blacks than for Whites (see Biernat & Fuegen, 2001; Biernat & Kobrynowicz, 1997; Biernat & Ma, 2005; Biernat, Ma, & Nario-Redmond, 2008). Thus, women and Blacks may have an easier time meeting minimal standards but a more difficult time meeting confirmatory standards with regard to attributes on which they are stereotyped as deficient. In one study, for example, women were *suspected* of being aggressive based on fewer behavioral examples of aggression, but were *confirmed* to be aggressive based on more behavioral examples, relative to men (Biernat et al., 2008). In another study, women were more likely to be placed on a short list for hiring (a minimum standard), but less likely to actually be hired, relative to men (a confirmatory standard of competence; Biernat & Fuegen, 2001).

These findings have emerged from research based on the “shifting standards model” (Biernat, Manis, & Nelson, 1991). The main premise of this model is that stereotypes operate as standards against which individual members of stereotyped groups are judged. Because standards differ for members of different groups, subjective evaluations of a man versus a woman or a Black versus a White are not directly comparable: “Competent” for a woman does not mean the same thing as “competent” for a man; “hardworking” does not mean the same thing when it is applied to a Black versus White employee. And because of stereotype-based low minimum standards for women and Blacks on such attributes, *more* evidence of competence or motivation is needed for them to *confirm* that they do, in fact, possess these qualities.

In this article, we further explore the implications of stereotype-based judgment standards for the evaluation and treatment of members of stereotyped groups. First, we reiterate the distinction between these standards when it comes to judgments of workplace *incompetence*. But more importantly, we then extend the analysis by examining how manipulations that subtly orient judges toward these differential standards affect *attention* to the performance of others and *decisions* about their future. The studies presented here focus on gender and race in the workplace and examine stereotypes regarding job-related incompetence.

Inferring Incompetence

When work performance is diminishing, what evidence is needed before a termination decision is made? Based on our distinction between minimum and confirmatory standards to infer competence (Biernat et al., 2008), we propose that members of groups stereotyped as competent (or in keeping with our earlier language, stereotyped as “deficient” in *incompetence*) should be held to lower *minimum* standards but higher *confirmatory* standards of incompetence, relative to members of groups stereotyped as relatively incompetent. In the present research, White men are stereotyped as deficient in

incompetence in many workplace domains, so minimum standards of incompetence should be lower and confirmatory standards higher for members of this group. That is, low expectations of incompetence should make it relatively easy for White men to meet minimum incompetence standards, but more evidence of incompetence should be tolerated from White men before a confirmation or diagnosis of incompetence is rendered.¹

Little research has addressed the inference of incompetence, but there is some suggestive evidence that a higher standard is used to confirm incompetence in those stereotyped as relatively competent. Linville and Jones (1980) demonstrated that a White law school applicant with weak credentials was judged more favorably than an identical Black applicant. Hass, Katz, Rizzo, and Bailey (1991) found that Whites who performed poorly in a team task were judged less negatively than Blacks who did the same. Similarly, Piacente, Penner, Hawkins, and Cohen (1974) found that male experimenters who were trained to act incompetently were judged less negatively than incompetent female experimenters. In short, the same poor performance elicits less *unfavorable* judgments of those stereotyped as deficient in incompetence.

One shortcoming of this past research is that it has assessed only subjective judgments of those displaying incompetence; it has not addressed standards for incompetence, particularly with regard to the distinction between minimum and confirmatory standards. In a relevant study from our lab, participants were asked to review the work record of a Black or White trainee whose on-the-job performance was clearly slipping (Fuegen & Biernat, 2010). Some participants were oriented toward a *probation* decision, which we conceptualized as a minimum standard of incompetence, and some toward a *firing* decision, which we conceptualized as a confirmatory standard of incompetence. Among those considering a Black target, probation and firing standards did not differ—a performance score of about 43% signaled that probation was in order and a score of 45% signaled that firing was in order. But among those considering the White target, a score of 50% was the signal for probation but a score of 36% signaled firing. That is, less evidence of bad performance was needed to trigger probation but more evidence of bad performance was needed to trigger firing for the White target.² Thus, we have some suggestive, though limited, evidence that group members stereotyped as deficient in incompetence (e.g., White men) may be held to lower minimum but higher confirmatory standards of job-related incompetence.

Effects of Standards on Attention to Incompetent Behavior

In the present studies, our goal was to more firmly establish the role of minimum and confirmatory evidentiary standards in inferences about workplace incompetence and to extend our analysis to how standards might affect attention to behavioral evidence and decisions about employees’ futures. Study

1 was designed to directly test the prediction that White men would be held to lower minimum but higher confirmatory standards of incompetence, relative to White women. Participants were asked to consider behavioral evidence relevant to workplace incompetence and to indicate the number of behaviors needed to meet minimum versus confirmatory standards (see Biernat et al., 2008).

Consistent with past research, we anticipate a statistical interaction between target category (male/female) and standard, with our primary prediction focusing on the group stereotyped as deficient on the attribute of interest (White men, stereotyped as deficient in incompetence): Minimum standards should be lower and confirmatory standards should be higher for White men than for Blacks and women. In other words, an effect of standard (minimum vs. confirmatory) should emerge in the case of White men, but not in the case of women. Although one might guess that minimum standards will always be lower than confirmatory standards, we suggest that “suspicion = confirmation” for group members stereotyped as having the trait in question, that is, stereotyped as incompetent (Biernat et al., 2008, p. 307). Of course this key predicted pattern could also mean, mathematically, that target category effects emerge within each level of standard—for example, men are held to lower minimum standards of incompetence but higher confirmatory standards of incompetence than women—but whether these simple effects are reliable is secondary to our main hypothesis.

More importantly, Studies 2 and 3 move beyond the artificial “minimum” and “confirmatory” labels and focus instead on how different evaluative mind-sets or tools may instantiate these standards. In earlier work, we considered how minimum standards of incompetence might be invoked by “probation” decisions but confirmatory standards by “termination” decisions (Fuegen & Biernat, 2009). In this research, we suggest that the *formality* of one’s evaluative focus or evaluative tools may invoke different standards as well. A “formal record” or “formal log” of an employee’s behavior may instantiate confirmatory standards; but informal “note taking” may make minimal standards more salient. In Studies 2 and 3, we address how minimum and confirmatory standards—invoked by informal versus formal evaluation tools—affect judges’ tendency to take note of poor performance of male versus female and Black versus White employees. These studies address two related questions: To what extent is negative performance of targets *attended to* among those considering minimal versus confirmatory standards of incompetence, and how do these standards affect decisions about termination?

In these studies, we exposed participants to behavioral evidence about a “trainee” that indicated mixed, though largely negative, performance during a probationary employment period. We asked participants to read through this performance file and to either take “informal notes” or keep a “formal employment log” regarding the trainee. We designed these

two types of records to differ in several ways. The “formal log” was described as reflecting a serious “permanent record,” whereas “informal notes” were for the participants’ eyes only. In this sense, the formal log represents a more consequential and serious record, one that we assumed would instantiate confirmatory standards for ability/inability and in which negative information would be harmful to the employee’s future. Informal notes, on the other hand, are less consequential in nature and might instantiate minimum standards, in that behaviors might be noted for their interest or surprise value.

We were particularly interested in judges’ tendency to record *negative* behavioral evidence of incompetence in these records. Based on our distinction between minimum and confirmatory standards of incompetence, we expected that participants would be less likely to record negative, incompetent behaviors in the *formal log* than in *informal notes* for men (in Study 2) and for Whites (Study 3). That is, the confirmatory standards invoked by a formal log will lead judges to notice and record fewer negative behaviors for groups stereotyped as competent (i.e., deficient in incompetence—viz., White men), but the minimum standards invoked by informal notes will lead judges to record *more* negative behaviors for this same group. Of course, lessened attention to negative information in the formal log condition could be based on higher confirmatory standards of incompetence for White men, *or* on straightforward hostile bias toward members of groups stereotyped as incompetent (women in masculine occupations; Blacks in white-collar occupations). But importantly, the predicted converse pattern among participants keeping “informal notes” suggests that differential standards, rather than hostile bias toward devalued groups, drive the attention to negative information. Where incompetence is less expected, it should be more readily noted—minimum standards should be met more quickly for White men.

Examining decision making about the target may also clarify whether lessened attention to negative information in the formal log condition is based on higher confirmatory standards of incompetence for White men (a cognitive account) or on hostile bias against women and Blacks (an affective account). That is, to what extent does the use of a formal log versus informal notes affect the decision to fire the employee? The cognitive (accessibility) account suggests that heightened attention to negative behaviors should lead to a negative decision (White males should be terminated more often in the informal notes than in the formal log condition), but an affective account suggests that despite the heightened attention to negative behaviors in the informal notes condition, White males will be terminated less often in this condition.

In Studies 2 and 3, we examine these alternative predictions by asking participants to both record information about a trainee’s performance and make a decision about his or her future in the position. In Study 2, targets are male versus

female trainees in either a masculine or feminine occupation. We predict that in the case of the masculine occupation, because men are stereotyped as deficient in incompetence, participants will record more incompetent behaviors for men when keeping “informal notes” than when keeping a “formal log.” We also predict that termination decisions will follow the same pattern: Men will be fired more often among participants keeping informal notes than those keeping a formal log. Study 2 also included a “feminine” job condition, designed to reverse the relevant stereotype and predicted pattern of results. In Study 3, we test comparable predictions for Black versus White male trainees. We expect that more incompetent behaviors will be recorded for White targets when “informal notes” are kept than when “formal logs” are kept. Termination decisions should also follow this pattern. We first turn to Study 1 to test the basic assumption that minimum standards of incompetence are lower but confirmatory standards are higher for men relative to women in a masculine occupational setting.

Study 1

Method

Participants were 102 undergraduates (38 women, 64 men; 83.3% White) at the University of Kansas who participated in an online study in exchange for course credit. Participants were randomly assigned to one of four conditions in a 2 (target sex) \times 2 (standard: minimum/confirmatory) between-subjects design.

All participants were told the study was about “the criteria people use when forming impressions or making decisions about others” and that we were focusing on standards for incompetence in the workplace. Following procedures used by Biernat et al. (2008), participants in the *minimum standards* condition read that we were interested in “the MINIMUM number of behaviors that are necessary to SUSPECT that a person may be incompetent”; the wording in the *confirmatory standards* condition was “the TOTAL number of behaviors that are necessary to CONFIRM that a person is incompetent” (emphasis original).

All participants were then asked to review a list of 10 behaviors and to check off “as many or as few behaviors” as a target employee would need to engage in to either “GIVE YOU SOME INKLING OR HINT that [the target] may be INCOMPETENT” or to “CONFIRM that [the target] is INCOMPETENT.” Crossed with this manipulation, participants were asked to imagine an employee named “Kenneth” or “Katherine,” who worked “at a high-tech computer hardware firm” (a masculine occupation), and were told that the behaviors indicate “some actions s/he might have engaged in on the job.” Participants were again instructed to read over the entire list of behaviors and then to check those that would be required to meet minimum or confirmatory standards of

incompetence. The key dependent measure was the number of behaviors (out of 10) checked.

The 10 behaviors, listed in Table 1, were selected from an earlier pretesting, in which 22 participants rated a set of 85 workplace behaviors on negativity/positivity. Those chosen for use in this study were rated in the 2.09 to 3.98 range on a 1–9 (*negative to positive*) scale. After the fact, we recognized that pretesting on the more specific dimension of *incompetence/competence* was in order. An independent sample of 17 judges rated the same 85 behaviors on a 1–9 (*indicates incompetence to indicates competence*) rating scale; means and standard deviations for the 10 items included in Study 1 appear in the last two columns of Table 1. The correlation between judgments of *positivity* and *competence* across the whole sample of 85 behaviors was .98; in this subset of 10, the correlation was .89 (even the standard deviations were highly correlated: .62 in the full sample and .79 in the Study 1 set). Nonetheless, one of the negative behaviors was judged close to the midpoint rather than the lower end of the competence scale (“Complained that his/her prior coworkers tended to be lazy”), and therefore we report the analyses below with and without this item included in the count.

Results

A Target Sex \times Standard ANOVA on number of behaviors (out of 10) checked revealed the predicted interaction, $F(1, 98) = 14.92, p < .001$. Minimum standards ($M = 4.00, SD = 2.03$) were lower than confirmatory standards ($M = 6.88, SD = 2.32, p < .0001$) for Kenneth ($d = 1.32$), but for Katherine the minimum-confirmatory difference was not significant ($M_s = 4.85$ and $4.38, SD_s = 1.82$ and $2.54, p > .46$). In addition, *confirmatory* standards of incompetence were lower for Katherine than Kenneth ($p < .001, d = 1.03$), and *minimum* standards of incompetence were nonsignificantly lower for Kenneth than Katherine ($p < .15, d = -0.44$). In short, the male target more easily triggered the *suspicion* of incompetence but he had to engage in more incompetent behaviors before *sealing* that judgment, relative to the female target. Dropping the behavior that did not clearly pretest as incompetent, the Target Sex \times Standard effect remained significant, $F(1, 98) = 13.11, p < .001$, as were the same three of four simple effects.³

Discussion

Study 1 supports the basic premise of our research about the distinction between minimum and confirmatory standards of incompetence. Men, stereotyped as competent (or deficient in *incompetence* relative to women), were held to lower minimum standards but higher confirmatory standards to diagnose incompetence in a masculine job. For women—stereotyped as relatively incompetent in a masculine domain—suspicion and confirmation invoked identical evidentiary requirements.

Table 1. Employee Behaviors Used in Study 1, Along With Pretested Negativity/Positivity and Incompetence/Competence Mean Ratings and Standard Deviations

Behavior	Positivity (negative–positive, 1–9)		Competence (incompetent– competent, 1–9)	
	M	SD	M	SD
Lost a file on a client.	2.12	1.27	2.06	1.34
Forgot about a previously scheduled appointment with a client.	2.16	1.17	2.18	1.13
Missed an important deadline.	2.50	1.53	2.59	1.77
Asked about getting extra vacation time.	3.82	1.50	3.35	1.80
Took home supplies, such as pens and envelopes, from the company.	2.79	1.45	2.71	1.72
Didn't review the training manual as the training supervisor instructed.	2.62	1.34	2.82	1.51
Asked the vice president about getting better parking privileges.	3.86	1.77	3.76	1.75
Took an important message for a coworker and then misplaced it.	2.48	1.46	3.06	1.52
Complained that his/her prior coworkers tended to be lazy.	3.98	1.57	4.76	1.52
Inadvertently gave confidential information to the competition.	2.09	1.52	2.47	1.70

This study documents that whether stereotypes about incompetence lead to stringency or leniency in evidentiary standards for male targets depends on the standard at hand.

These findings also set the stage for further inquiry about the consequences of differential standards for workplace outcomes. Studies 2 and 3 examine whether these standards—instigated in subtle ways, through the use of different evaluative tools—affect judges' attentiveness to negative behavior in the workplace and, in turn, their decisions about the employee's future. In these studies, participants are asked to review the performance of an employee, keeping track of notable behavior in either their "informal notes" or a "formal employment log." We test our assumption that the formal log instantiates confirmatory standards, whereas informal notes instantiate minimum standards of incompetence through pretesting. We focus on the number of negative behaviors recorded in these formats and predict that for group members stereotyped as competent (or as deficient in incompetence), minimum standards will be lower (more negative behaviors recorded in informal notes) but confirmatory standards will be higher (fewer negative behaviors recorded in formal notes), relative to those stereotyped as incompetent. Decisions to terminate or not are expected to follow from the attention to and accessibility of negative information.

Study 2

Method

Participants. Participants were 83 undergraduates (43 women, 40 men; 88% White) at the University of Kansas who participated in exchange for course credit.

Procedure. Participants were conducted through the experiment individually. They were told they would be role-playing the position of Director of a training program who was in

charge of evaluating trainees at a large corporation. Participants were told they would evaluate one trainee who had been working at the company for 3 months, and they were provided with a resume and a description of the trainee's job. The job description was identical to one used in previous research; it described a generic office position, except for half the participants the job title was masculine ("Executive Chief of Staff") and for the other half it was feminine ("Executive Secretary"; see Biernat & Fuegen, 2001; Biernat & Kobrynowicz, 1997). All participants viewed the same resume, except the name indicated the trainee was either male (Kenneth Anderson) or female (Katherine Anderson).

Participants were then given a brief summary of the trainee's performance, which contained 16 examples of behaviors the trainee had engaged in during the 3-month program. These behaviors are displayed in Table 2, along with their pretested *positivity* and *competence* values (see description of pretesting in Study 1). We selected eight negative/incompetent behaviors, four positive/competent behaviors, and four behaviors that pretested as neutral on these dimensions (the positivity–competence correlation was .98 in this set of behaviors).

Along with the summary of the trainee's performance, participants received a single sheet of paper labeled either "Your notes to yourself" or "Performance log." Each sheet contained several lines for recording information. The instructions for these two logs were as follows:

Your notes to yourself: Please record any notable information regarding the performance of the trainee using this sheet to take informal notes for yourself. This sheet is to help you in giving feedback to the trainee, so you can jot down anything that might be useful to you as you try to help in training the individual. These notes are for your eyes only. They are an informal way of keeping track of your thoughts about the trainee's

Table 2. Employee Behaviors Used in Studies 2 and 3, Along With Pretested Negativity/Positivity and Incompetence/Competence Mean Ratings and Standard Deviations

Behavior	Positivity (negative– positive, 1–9)		Competence (incompetent– competent, 1–9)	
	M	SD	M	SD
Negative/incompetent behaviors				
Trainee erased an important document from the computer.	2.07	1.37	2.82	1.94
During a training session, trainee mentioned that the reason s/he chose business administration as a college major was because his/her parents wouldn't pay for anything else.	2.71	1.66	3.00	1.32
Trainee didn't review the training manual as the training supervisor instructed.	2.62	1.34	2.82	1.51
Trainee made several mistakes because s/he refused to ask for help.	2.14	1.27	2.00	1.12
Trainee forgot about a previously scheduled appointment with a client.	2.16	1.17	2.18	1.13
A secretary asked the trainee how to be more effective. Trainee's response was that the secretary should be responsible for coming up with ideas for how to be used more effectively.	3.05	1.80	3.35	2.26
Trainee took an important message for a coworker and then misplaced it.	2.48	1.46	3.06	1.52
Trainee misled the personnel manager regarding his/her knowledge of a particular computer program.	2.80	1.46	2.88	1.41
Positive/competent behaviors				
The trainee manager noted that the trainee uses appropriate interpersonal styles when dealing with client phone calls.	6.96	2.09	7.29	2.05
Trainee said that most of his coworkers are dependable.	6.59	1.72	6.53	1.62
Trainee was given a series of projects to complete, the first of which was finished on time.	7.45	2.15	7.24	2.28
Trainee looked into the possibility that the company would pay for classes in time management.	6.82	1.69	7.59	1.12
Neutral behaviors				
Trainee commented that the company morale wasn't very positive.	4.45	1.40	5.29	1.72
A client called to complain that the trainee was rude. Trainee responded that the client was rude and that he did his best to be helpful.	4.59	1.79	4.59	2.06
A manager noted that the trainee could be better at being a team player, but instead works better at one-on-one relationships.	5.62	1.33	5.29	1.26
Trainee told a coworker that he often feels stressed due to the daily workload.	5.07	1.26	4.94	1.56

progress. They won't be used for promotion or salary purposes. Your notes are only for your use in working with the trainee in the future. These notes do not have any serious implications for the trainee.

Performance log: Please record any notable information to yourself regarding the performance of the trainee using this formal performance log. This log is to help you in giving feedback to the trainee, so you can record anything that might be useful to you as you try to help in training the individual. But this log is not for your eyes only. It is a formal way of keeping track of the trainee's progress. It will be used for promotion and salary purposes and will become part of the trainee's permanent employment file, for all managers and executives to review when considering the future of this trainee in the company. Your goal with these notes is to track the trainee's performance.

Negative information in the log may have serious implications for the trainee.

Participants were told to read through the summary of the trainee's performance while recording any information they believed should go into in their informal notes/performance log. Participants were then asked to indicate whether the company should train the target for another 6 months, give him/her a job with the company, or terminate the trainee. At this point, participants were debriefed, thanked, and excused.

In addition to the termination decision, the main dependent variable of interest was the number of negative/incompetent behaviors (of 8 included in the original description) participants recorded on the informal notes/performance log. The number of positive (of 4) and neutral/ambiguous behaviors (of 4) recorded was also assessed. Two independent judges (blind to condition) read each participant's entries and

categorized them as positive, negative, or neutral; these judges were in strong agreement, with correlations between judges ranging from .68 for neutral behaviors to .84 for positive and .86 for negative behaviors. We averaged across the two coders to create the number of negative, positive, and neutral behaviors recorded for each trainee.

Pretest of record/evaluative tool manipulation. To evaluate the viability of our record type manipulation, we provided a separate sample of 47 undergraduates with a brief version of the manipulations described above (performance log or informal notes) and asked them to indicate the extent to which negative behaviors recorded in these logs would (a) be “helpful or hurtful to the employee,” (b) “have serious implications for the employee,” (c) indicate that “the employee’s future prospects” were good, and (d) “indicate that the manager was interested in” the employee’s future. Each item was answered on a scale from 1 to 7, with end points appropriately labeled. Indicated by *t* tests, negative behaviors recorded in the formal log relative to informal notes were perceived as (a) less helpful to the employee, $M_s = 2.71$ and 4.26 for formal and informal log, respectively, $t(45) = 2.96$, $p < .01$, (b) more serious in implications, $M_s = 5.17$ and 4.13 , $t(45) = 2.54$, $p < .02$, (c) more likely to indicate the employee was headed toward failure, $M_s = 6.25$ and 5.57 , $t(45) = 2.03$, $p < .05$, and (d) somewhat less likely to indicate interest in the employee’s future, $M_s = 2.79$ and 3.74 , $t(45) = 1.70$, $p < .10$. We did not ask whether the record indicated the use of minimum versus confirmatory standards, as past attempts at such questions suggest that respondents are unable to provide meaningful answers. However, we did ask a separate sample of 14 students to indicate “how negative a trainee’s behavior would have to be before it was recorded” in the formal log/informal notes. Participants indicated that behavior would have to be significantly more negative before it made its way into the formal performance log ($M = 5.17$) than the informal notes ($M = 4.14$), $t(12) = 2.38$, $p < .05$.

Results

Manipulation checks. Two participants incorrectly identified the sex of the applicant they reviewed and were omitted from subsequent analyses, leaving a sample of 42 females and 39 males. As a check on the log manipulation, participants were asked to indicate whether their recordings would be seen “just by me” or “by others.” Roughly 80% of participants in each condition made the correct identification, $\chi^2(1 df, N = 81) = 23.48$, $p < .0001$, but all were maintained in the analyses that follow (results did not change when those in error were excluded).

Recording negative behaviors. The number of negative behaviors participants recorded (of eight possible) was submitted to a Trainee Sex \times Record Type \times Position \times Participant Sex ANOVA. A main effect of Record Type indicated that more negative behaviors were recorded in informal notes ($M = 5.15$,

$SD = 2.07$) than in the formal log ($M = 3.95$, $SD = 1.84$), $F(1, 65) = 7.64$, $p < .01$. Evidence of ingroup favoritism was also noted in the Trainee Sex \times Participant Sex interaction, $F(1, 65) = 10.64$, $p < .01$. Female participants recorded more negative behaviors for Kenneth ($M = 5.07$, $SD = 1.95$) than Katherine ($M = 3.75$, $SD = 1.71$; simple effect $p < .05$, $d = 0.72$), whereas male participants recorded more negative behaviors for Katherine ($M = 5.45$, $SD = 2.02$) than Kenneth ($M = 3.82$, $SD = 2.02$; simple effect $p < .02$, $d = -0.81$).

Independent of these effects was the significant Trainee Sex \times Record Type \times Position interaction, $F(1, 65) = 5.47$, $p < .05$. The relevant means are displayed in Table 3. Looking first at the left side of the table (the Chief of Staff position), the Trainee Sex \times Record Type interaction was significant, $F(1, 65) = 7.39$, $p < .01$, and the pattern of means was precisely as predicted: Among those considering Kenneth, more negative behaviors were recorded in the informal notes than formal log condition ($p < .01$, $d = 1.60$), but record type had no effect on number of negative behaviors recorded for Katherine, $p > .35$. In addition, slightly more negative behaviors were recorded for Kenneth than for Katherine in the informal notes condition ($p < .07$, $d = -0.69$), but slightly fewer negative behaviors were recorded for Kenneth than Katherine in the formal log ($p < .06$, $d = 1.02$). An additional analysis of *percentage* of negative behaviors recorded (out of all recorded behaviors) produced the same pattern of effects: The Trainee Sex \times Record Type interaction was significant for the Chief of Staff condition, $F(1, 65) = 5.84$, $p < .02$.

The means for the secretarial position are displayed on the right side of Table 3. The Trainee Sex \times Record Type interaction was not significant ($F < 1$, $p > .55$). Participants only recorded more negative behaviors in informal notes than the formal log, $F(1, 65) = 5.40$, $p < .05$.⁴

Termination decisions. Participants indicated whether they thought the trainee should be retained in the training program for another 6 months, given a job, or terminated. Only 1 participant indicated the employee should be given a job, so this response was recoded along with the “keep in training” responses as a “maintain” response. We then transformed these responses into a 0–1 (maintain–terminate) variable and submitted this to a Trainee Sex \times Record Type \times Position \times Participant Sex ANOVA. No effects were significant, including the predicted Trainee Sex \times Position \times Record Type interaction, $F(1, 64) = 1.65$, $p < .21$. Given this null effect, it was statistically inappropriate to consider the follow-up two-way interactions, though the pattern of means was consistent with predictions in the Chief of Staff condition. Low power may have contributed to this problem; we examine termination decisions more fully in the larger sample of Study 3.

A final question was whether the number of negative behaviors recorded was related to the decision to terminate trainees. Overall, the correlation between these variables was modest but positive ($r = .28$, $p < .05$; $n = 81$); the more negative

Table 3. Number of Negative Behaviors Recorded by Trainee Sex, Position, and Record Type, Study 2

Trainee	Position and record type							
	Chief of Staff				Secretary			
	Informal notes		Formal log		Informal notes		Formal log	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Katherine	4.22 ^b	2.40	4.90 ^c	2.02	5.45	2.15	3.70	1.27
Kenneth	5.70 ^{ab}	1.87	3.25 ^{ac}	1.09	5.10	1.85	3.95	2.42

Note: Means with matching superscripts differed as follows (all two-tailed tests):

a. $p < .01$.

b. $p < .07$.

c. $p < .06$.

behaviors recorded, the more likely the decision to terminate the trainee.

Discussion

Attention to negative information was affected by the evaluation tool participants used, at least in the masculine (Chief of Staff) condition. As predicted, participants recorded more negative behaviors for Kenneth in the *informal log* than in their *formal notes*, but this evaluation tool manipulation had no effect on attention to negative behavior for Katherine, stereotyped as relatively incompetent in this domain. We suggest that these patterns are because of the instantiation of different standards of incompetence—minimum standards in the case of informal notes and confirmatory standards in the case of the formal log—that affect judges' orientation to behavioral evidence. In the case of the formal log, members of groups stereotyped as deficient in *incompetence*—men in this case—appeared to be held to higher evidentiary standards: Fewer negative behaviors were recorded in the “permanent record.” Importantly, the predicted converse pattern was found in the case of informal notes: Negative behaviors were more notable and more likely to be recorded for men. This finding suggests that attention to negative information was driven by differential standards rather than hostile bias toward devalued groups.

This pattern replicates Study 1, but importantly, it was based on a more realistic and more *subtle* manipulation. Employers probably rarely think in terms of “minimum” versus “confirmatory” standards, and they rarely check off behaviors that would provide evidence that a standard has been met. By providing evaluative tools in Study 2 (“notes” vs. “formal log”), and by asking participants to record whatever they desired using these tools, we better captured the way standards may be activated in the real world and the subtle way in which these standards drive attention to evidence of incompetence for women versus men.

We know from our pretesting that formal logs and informal notes are perceived to differ in a number of ways, one of which is the *seriousness* of the former and its apparent

instantiation of a higher standard (i.e., behaviors need to be “more negative” before being recorded here). But another difference was that recording negative behaviors in informal notes was judged to be potentially more helpful to the employee and less likely to indicate he or she was heading toward failure. Thus, one might argue that the greater recording of the male trainee's negative behaviors in informal notes was motivated not by standard shifts but by the desire to help the male (or not help the female). However, the significant positive correlation between negative behaviors recorded and termination likelihood makes it unlikely that the recording of negative behaviors was motivated by helpfulness. Instead, we suggest that the activation of different standards led to a differential focus of attention on negative behavior that, in turn, prompted firing. But power was low in this study (with *ns* at roughly 10 per cell), and we were unable to reliably detect condition differences in termination, or to consider any meditational analysis. We return to this issue in Study 3.

Our Study 2 findings were generally consistent with predictions in the Chief of Staff condition, but we did not find evidence of a reverse pattern when participants considered trainees in the feminine, “Executive Secretary” position. Although stereotypes suggest that women will be more competent than men in such a position (that women are deficient on *incompetence*), there were no significant differences in termination rates, or in number of negative behaviors recorded based on trainee sex. We can only note that our own past research on standard setting has documented stereotyping effects more strongly in the case of masculine than feminine jobs, or has found a pattern across job types indicating greater competence perceptions of men (e.g., see Biernat & Fuegen, 2001; Biernat & Kobrynowicz, 1997). Perhaps the use of the modifier “Executive” to describe the “secretarial” position was somewhat male stereotypic, counteracting the femininity of the “secretary” role. Consistent with this possibility, when we asked participants (in the manipulation checks section) to indicate how “high or low in status” they perceived the position to be, the Executive Chief of Staff and Executive Secretary positions were perceived as equal and

relatively high in status (on a 7-point scale, Chief of Staff $M = 5.33$, Secretary = 5.46, $F < 1$).

Another possibility is that the male–competence link is diffuse enough that it extends across many occupational settings. In describing gender-based double standards of competence, Foschi (1992) notes that “if the possibility of a relationship between sex and task competence is left open, the diffuse nature of the former characteristic will still enable the formation of higher expectations for the man than for the woman” (p. 185). In short, although we assume that the secretarial position was feminine, presumably invoking some expectation of female competence, this was counteracted by a general tendency to expect competence in men. Though admittedly post hoc, this may nonetheless explain the null effects of trainee sex in the “Executive Secretary” condition. Future work will clearly be necessary to determine whether the patterns produced here can be reversed in cases where women are stereotyped as competent (deficient in incompetence) relative to men.

For the most part, the Study 2 results were not affected by participant sex: Male and female participants showed the same pattern of judgments in all conditions, consistent with the notion that gender stereotypes are shared and acted on by both men and women (e.g., see Blair & Banaji, 1996; Eagly & Mladinic, 1989; Rudman & Glick, 2001). One exception to the lack of participant sex effects was the evidence of ingroup bias we found in recording of negative behaviors: Women were more likely to record negative behaviors for Kenneth than Katherine, and men were more likely to record negative behaviors for Katherine than Kenneth. This finding was independent of the trainee sex by record type interaction; participant sex did not moderate that key effect. Nonetheless, the finding suggests that other factors, including ingroup bias, may be operative in an employee evaluation setting (Fuegen & Endicott, 2009). That this ingroup favoritism did not extend to firing, however, indicates that the effect was not strong or consistent.

In Study 3, we leave behind the issue of gender and focus instead on the role of trainee race in attention to negative behavioral evidence and termination decisions. Study 3 uses only the “Executive Chief of Staff” condition of Study 2 and exposes participants to either a Black or White male performing poorly in the trainee program. To address concerns about statistical power in Study 2, we substantially increased the sample size, which will be particularly important for detecting effects on termination and for testing mediation (via attention to negative behaviors) of a predicted race by record type interaction on termination decisions.

Study 3

Method

Study 3 was a replication of Study 2 that focused on race rather than gender, comparing responses to a Black versus

White male employee whose performance during the internship was poor. Participants were 152 undergraduates (88 women, 64 men; 97.4% White) who received credit in their Introductory Psychology courses. The method and measures were identical in all respects to Study 2, with the exception that the applicant was named either “Denzel Carter” or “David Winkler,” names that were pretested to imply Black or White racial identification, and only the Executive Chief of Staff description was used. One judge coded all behaviors participants recorded in their formal log or informal notes, and a second independent judge coded a subset (roughly half). Agreement ranged from .81 to .91 for the three behavior valences. Thus, the codings of the first judge were used in all analyses.

Results and Discussion

Manipulation checks. One participant incorrectly indicated the target was female and was dropped from the analyses, leaving 87 women and 64 men in the sample. As a race manipulation check, we asked an open-ended question at the end of the study: “Although we didn’t explicitly give you this information, what was your best guess about the trainee’s race?” Of those exposed to “David Winkler,” 90.1% guessed he was White (1 participant guessed “Asian,” and 6 guessed “Black”). But of those exposed to “Denzel Carter,” only 57.5% correctly guessed he was Black (the rest guessed he was White). This difference supporting the race manipulation in race attribution was clearly significant, $\chi^2(2 \text{ df}, N = 151) = 40.56, p < .0001$, but the fairly common attribution of “White” to Denzel Carter raises some concerns. We suspect, however, that many participants were simply reluctant to admit they had paid attention to race and wrote “White” as the default. This may be analogous to recent findings suggesting that Whites are relatively unlikely to acknowledge race in interracial situations (Apfelbaum, Sommers, & Norton, 2008). We maintained all participants in the analyses reported below but also conducted additional analyses using the subset of participants who correctly identified the target’s race. The substantive results did not change in this subsample, as indicated in notes 5 and 7. The log manipulation check indicated that roughly 86% of the sample correctly identified whether their recordings would be seen “just by me” or “by others,” $\chi^2(1 \text{ df}, N = 150) = 77.17, p < .0001$, and all were maintained in the analyses.

Recording negative behaviors. We predicted that participants would record more negative behaviors in the informal notes than in the formal log for David, the White target. The number of negative behaviors participants recorded (of 8 possible) was submitted to a Trainee Race \times Record Type \times Participant Sex ANOVA. The only significant effect was the Trainee Race \times Record Type interaction, $F(1, 141) = 5.25, p < .05$. As can be seen in the left-hand panel of Table 4, the pattern of means was as predicted: Among those participants considering the White target, more negative behaviors were recorded

Table 4. Number of Negative Behaviors Recorded and Termination Decisions by Trainee Race and Record Type, Study 3

Trainee	Dependent variable and record type							
	# of negative behaviors recorded				Proportion termination decisions			
	Informal notes		Formal log		Informal notes		Formal log	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Denzel	4.40	3.05	4.83 ^b	2.56	0.45	0.50	0.54 ^d	0.51
David	4.88 ^a	2.11	3.00 ^{ab}	2.57	0.51 ^c	0.51	0.32 ^{cd}	0.48

Note: Means with matching superscripts differed as follows (all two-tailed tests):

a. $p < .05$.

b. $p < .01$.

c. $p < .06$.

d. $p < .08$.

in the informal notes than in the formal log, $p < .05$, $d = 0.80$. This difference was not significant for the Black target, for whom suspicion of incompetence may equal confirmation. This interaction was also driven by a significant White–Black difference in the formal log condition, $p < .01$, $d = 0.71$; the target race effect in the informal notes condition was not significant, $p > .25$.

In an additional analysis, the percentage of negative behaviors recorded (out of all behaviors recorded) was also submitted to a Trainee Race \times Record Type \times Participant Sex ANOVA. Again, the Trainee Race \times Record Type interaction was significant, $F(1, 131) = 10.04$, $p < .01$.⁵ For David, 46% of behaviors recorded in the formal log were negative ($SD = 0.27$), compared to 67% in the informal notes condition ($SD = 0.25$, $p < .01$, $d = 0.83$). For Denzel, negative behaviors were recorded 57% of the time in the formal log ($SD = 0.21$) and 47% of the time in informal notes ($SD = 0.25$), a marginally significant difference ($p < .10$). The race difference was significant in the informal notes condition ($p < .01$, $d = 0.80$).⁶

Termination decisions. Participants indicated whether they thought the trainee should be retained in the training program for another 6 months, given a job, or terminated. In this study, 6 participants indicated the employee should be given a job (4 in the Black condition, 2 in the White), and this response was again recoded along with the “keep in training” responses as a “retain” response (scored as 0), to be compared with “terminate” responses (scored as 1). These were submitted to a Trainee Race \times Record Type \times Participant Sex ANOVA. As can be seen in the right panel of Table 4, the termination decisions followed the predicted pattern and mirrored the results for recordings of negative behavior; the Trainee Race \times Record Type interaction was significant, $F(1, 142) = 4.32$, $p < .05$. Among those participants considering the White trainee, termination was recommended more often in the informal notes than formal log condition ($p < .055$, $d = 0.39$). Record type did not affect termination decisions for the Black trainee ($p > .25$). In addition, the race difference

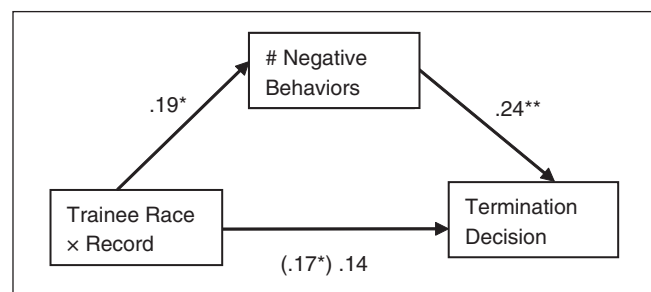


Figure 1. Trainee Race \times Record Type effect on termination mediated by number of negative behaviors recorded, Study 3

was marginally significant in the formal log condition ($p < .08$, $d = 0.43$).⁷

Finally, as in Study 2, we examined whether the number of negative behaviors recorded was related to the decision to terminate trainees. Overall, the correlation between these variables was modest but positive ($r = .24$, $n = 148$, $p < .01$). The more negative behaviors recorded, the more likely the decision to terminate the trainee. We next tested a mediational model, in which number of negative behaviors recorded mediated the effect of the Trainee Race \times Record Type on termination decisions (Figure 1). Partial mediation was evident; the interaction effect on termination was reduced to nonsignificance ($p < .11$) when negative behaviors were included in the regression equation, and a Sobel test was marginally significant ($z = 1.80$, $p < .075$).

General Discussion

Gender and racial stereotypes may play a role in the inference of workplace incompetence. Past research has examined how race and gender affect *evaluations* of poor performers, but much less is known about the evidentiary standards used to infer incompetence, to attend to evidence regarding incompetence, and to reach decisions about job termination. In the

three studies reported here, we distinguished between minimum and confirmatory standards and predicted that for groups stereotyped as deficient on an attribute, minimum standards are lower but confirmatory standards are higher, relative to members of groups stereotyped as having the attribute in question. In these studies, we assume White men are stereotyped as deficient in *incompetence* and therefore minimum standards are lower but confirmatory standards are higher for inferring incompetence in White men.

Study 1 was designed to directly measure how much evidence of incompetence was necessary for perceivers to meet minimum versus confirmatory standards for male versus female targets. The results were completely consistent with predictions—minimum standards were lower but confirmatory standards were higher for “Kenneth” but not for “Katherine.” We suggest that because men’s incompetence is unexpected, one is alerted to it more quickly (based on less evidence of incompetent behavior) but is also confident of it more slowly (see also Biernat et al., 2008).

Studies 2 and 3 extended this analysis by examining judges’ attention to negative performance information, as indicated by what was recorded in a “formal performance log” versus “informal notes.” We suggest that a formal log instantiates a confirmatory standard of incompetence—when information will appear in a “permanent record,” negative behavior should be recorded to the extent that it confirms incompetence. Indeed, pretesting indicated that behavior would need to be more negative before it was included in a formal log rather than informal notes and that recording negative behaviors in the formal log had more serious implications for the employee. Informal notes, on the other hand, may instantiate a minimum standard of incompetence. To the extent that one is merely “jotting down” notes to oneself, confirmation of incompetence is not required; instead, one may be sensitive to violations of expectations.

Based on stereotypes of White men as relatively competent (or deficient in *incompetence*) we predicted that those considering a White male trainee would record *more* negative behaviors in informal notes than in the formal log (in the Chief of Staff position). Studies 2 and 3 supported this pattern, in terms of raw numbers of negative/incompetent behaviors recorded and percentage of all recorded comments that were negative. We suggest that because incompetence is less expected in men (minimum standards are lower) negative behavioral evidence is noted more readily when using an informal evaluation tool. At the same time, confirmatory standards of incompetence are high, and therefore evidence of incompetency may be less likely to make its way into a formal performance log.

Studies 2 and 3 also indicated that for the target stereotyped as incompetent (the female Chief of Staff trainee in Study 2, the Black trainee in Study 3), roughly *equal* numbers of negative behaviors were recorded regardless of the evaluation tool. Similarly, in Study 1, the effect of standard

(minimum vs. confirmatory) was significant only in the case of the male, but not the female, target. This pattern suggests that for members of groups stereotyped as possessing a given trait, suspecting and confirming amount to the same thing. It is only when one is *not* expected to be incompetent that minimum standards are lower than confirmatory standards or that attention to evidence of incompetence varies depending on the evaluative tool being used.

Our primary hypotheses focused on the predicted effect of standard or evaluative tool in the case of White men but not White women or Black men. But it is worth noting that this pattern of statistical interaction may mathematically produce *target* effects within standard condition: Confirmatory standards of incompetence may be higher for White men than for women and Blacks, and minimum standards may be lower for White men than for women and Blacks. Across the three studies, target effects were most striking in the case of *confirmatory* standards and the *formal log* condition (simple effect $ps < .001, .06, \text{ and } .01$ in Studies 1, 2, and 3, respectively) rather than *minimum* standards and *informal notes* (comparable $ps > .08$). This suggests that evidentiary *stringency* toward White men when it comes to confirming incompetence (more evidence of incompetence required; fewer negative behaviors noted in the log) is more robust than the leniency (less evidence of incompetence required) and heightened attentiveness to negativity when minimum standards are invoked.

Standards, Evaluation Tools, and Firing

Although we have focused on the differential standards invoked by the formal log/informal notes manipulation in Studies 2 and 3, these records certainly differ in other ways as well. Pretesting data indicated that perceivers viewed negative behaviors recorded in informal notes as potentially more helpful and less likely to indicate an employee was headed toward failure, than negative behaviors recorded in a formal log. Thus, one might attribute the greater frequency of negative behaviors recorded for men and Whites in informal notes to a desire to be helpful to these trainees, perhaps by aiding them in changing their ways. However, we found that termination decisions followed the same pattern as the negative behavior recordings, with White men in Study 3 fired less often when the formal log was used than when informal notes were used (the comparable pattern in Study 2 was in the same direction, though nonsignificant). In addition, the greater the number of negative behaviors recorded, the more likely participants were to terminate the trainee. These patterns suggest that that helpfulness to the White male trainees was unlikely to be a primary motive when negative behaviors were recorded in informal notes.

Instead, we suggest that the different types of evaluation tools invoked different standards which, in turn, led to the recording of differential numbers of negative behaviors,

depending on target category. The accessibility of negative information, in turn, produced a tendency to fire.

We must acknowledge that we do not have direct evidence that the record manipulation invoked different standards. Our past research suggests that reporting on standards is difficult, and being able to articulate a difference between minimum and confirmatory standards may be even more so. But the pretesting data provide some evidence supportive of this premise, and our findings regarding the record type manipulation map onto those of Study 1, in which minimum and confirmatory standards were explicitly manipulated. In Study 3, where power was sufficient to test the mediational model, there was evidence of at least partial mediation (via number of negative behaviors recorded) of the race by record type effect on termination decisions (see Figure 1).

Future Directions

Our research fell short in documenting the context-specific nature of gender stereotypes. In Study 2, we expected that an “Executive Secretary” position would imply femininity, and that women would be stereotyped as competent, producing a reverse pattern of the target sex by record type effect reported in the left-hand panel of Table 3. Though the means were in the predicted direction, the results are better characterized as null. Perhaps our instantiation of the feminine position was weak, or perhaps men are afforded a diffuse air of competence across a number of work settings (Foschi, 1992). Because stereotyped expectations are at the heart of our phenomenon, future research should be able to identify contexts in which women and Blacks, rather than White men, are held to lower minimum but higher confirmatory standards of incompetence.

Future research could also consider more fine-grained analyses of gender and racial stereotypes. We focused on *White* targets when studying gender effects (Studies 1 and 2) and *male* targets when studying race effects (Study 3). Whether the same gender patterns hold for Black as for White targets, and the same race patterns for female as for male targets, remains to be seen. To the extent that stereotypes about competence/incompetence are relevant, we expect minimum and confirmatory standards should differ for members of contrasting target groups. We also suspect that motivational concerns (such as ingroup bias, evident in Study 2) as well as group stereotypes play a role in the consideration of behavioral evidence relevant to incompetence. Further research is needed to address this issue, perhaps by asking judges about their motivations, or tapping online thoughts during the trainee review period using think-aloud procedures. In addition, it would be useful to examine minimum and confirmatory standards using a within-subjects design, as decision making about employees may naturally proceed from suspicion to confirmation, or from “notes” to the “permanent record.”

Such an approach may enhance the external validity of our findings.

In the studies reported here, the behaviors we used to instantiate incompetence were also judged as highly negative by pretest judges, with an overall correlation between behavioral positivity and competence (as judged by independent raters) of .98. Indeed, we have used the terms “incompetent” and “negative” roughly interchangeably throughout. But this raises the question of whether evidentiary standards of incompetence and evaluative tools affect attention to *incompetent* behaviors in particular or to *negative* behaviors more generally (including those indicating immorality, lack of warmth, etc.). In the present studies, we did not distinguish these possibilities, as all of the “negative” behaviors were also considered incompetent (with the exception of 1 behavior in Study 1). Further work could do so by including more differentiated negative behavioral episodes in the employee’s history. Our standards-based predictions focus on attention to incompetence per se, but whether a kind of “halo” effect emerges when considering behavioral evidence—as it often does in employee evaluation more generally (Saal, Downey, & Lahey, 1980)—is worth investigating.

Conclusions

Judging incompetence is an important though understudied process. In academic, workplace, and legal settings, the standards used to decide that an individual is incompetent, and the standards against which one begins noticing incompetence, matter for real-work outcomes including dismissals, demotions, and verdicts. The present research has suggested that category membership affects these standards, such that groups stereotyped as relatively competent (or deficient in incompetence) are held to lower *minimum* standards of incompetence (initial suspicion of incompetence is triggered sooner, as it is less expected), but higher *confirmatory* standards of incompetence (more evidence of incompetence is required before it is confirmed). The present studies revealed these tendencies directly in the setting of evidentiary standards and indirectly in the attention to negative behaviors prompted by the use of informal notes (minimum standard) or a formal performance log (confirmatory standard). Studies 2 and 3 also revealed that firing patterns generally followed from the attention to negative information. Additional work is needed to both delve more fully into the process involved (perhaps by gathering direct evidence of the role of differential standards, in combination with motivational factors) and examine the consequences of this pattern for both the target and the decision maker.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding

The authors received the following financial support for the research and/or authorship of this article: This research was supported, in part, by National Institute of Mental Health Grant 48844 awarded to the first author. Additionally, writing of this article was facilitated by NSF grant number 0820705.

Notes

1. The prediction regarding high confirmatory standards of incompetence is also articulated by Foschi (2000), who has argued that when it comes to inferring *lack* of ability, “the higher the status [of the actor], the more convincing the demonstration of incompetence will have to be” (p. 25; also see Foschi, 1989). This perspective is silent, however, on the minimum standards prediction.
2. We are aware of only one other line of research that examined category (gender) bias in firing decisions (Levin, Rouwenhorst, & Trisko, 2005). In two studies, these authors found that male employees were more likely to be fired when participants had to choose among 36 employees (18 male and 18 female) for downsizing, though there was no sex bias in the initial possible pool of candidates to fire. The authors suggest that the effect in the final firing choice was driven by emotion, “a more sympathetic attitude toward women” (p. 803).
3. A reviewer expressed concern that the “took home supplies, such as pens and envelopes, from the company” behavior was also not particularly relevant to incompetence. Though judge ratings indicated otherwise, we also analyzed the data with this behavior removed from the count. This interaction was again significant, $F(1, 98) = 11.29, p < .01$, as were the simple effects.
4. Participants recorded very few neutral ($M = 1.36$) or positive ($M = 1.16$) behaviors in their records. ANOVAs computed on these variables revealed no statistically significant effects for neutral behaviors, and only a theoretically uninteresting, marginally significant Record Type \times Position interaction for positive behaviors, $F(1, 65) = 3.92, p < .06$. In short, effects of trainee sex were detected only for number of negative behaviors recorded; a Behavior Valence (negative, neutral, positive) \times Trainee Sex \times Record Type \times Position \times Participant Sex mixed-model ANOVA (with repeated measures on the first factor) confirmed that the predicted three-way interaction among trainee sex, record type, and position was moderated by behavior valence, $F(2, 130) = 4.65, p < .02$.
5. In the subsample of those who correctly identified the target’s race, the Trainee Race \times Record Type interaction remained significant for both number of negative behaviors, $F(1, 100) = 5.38, p < .05$, and percentage of negative behaviors recorded, $F(1, 92) = 8.00, p < .01$.
6. As in Study 2, relatively few positive ($M = 1.30$) and neutral behaviors ($M = 1.74$) were recorded overall, and analyses of these data revealed no theoretically meaningful effects. A Behavior Valence \times Trainee Race \times Record Type \times Participant

Sex mixed-model ANOVA with repeated measures on the first factor (negative, positive, and neutral behavior recordings) produced a significant Behavior Valence \times Trainee Race \times Record Type interaction, $F(2, 282) = 6.74, p < .01$, further indicating that the predicted race pattern in informal notes versus formal performance log was specific to the recording of negative behaviors.

7. Using only the subsample of participants who correctly identified the target’s race, this interaction was marginally significant, $F(1, 102) = 3.66, p < .06$. A logistic regression also supported the reliable interaction effect on termination decision, $\chi^2(1 df, N = 151) = 4.63, p < .05$.

References

- Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology, 95*, 918-932.
- Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58*, 1019-1027.
- Biernat, M., & Fuegen, K. (2001). Shifting standards and the evaluation of competence: Complexity in gender-based judgment and decision making. *Journal of Social Issues, 57*, 707-724.
- Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology, 72*, 544-557.
- Biernat, M., & Ma, J. E. (2005). Stereotypes and the confirmability of trait concepts. *Personality and Social Psychology Bulletin, 31*, 483-495.
- Biernat, M., Ma, J. E., & Nario-Redmond, M. R. (2008). Standards to suspect and diagnose stereotypical traits. *Social Cognition, 26*, 288-313.
- Biernat, M., & Manis, M. (2007). Stereotypes and shifting standards: Assimilation and contrast in social judgment. In D. Stapel & J. Suls (Eds.), *Assimilation and contrast in psychology* (pp. 75-97). New York, NY: Psychology Press.
- Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology, 60*, 485-499.
- Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142-1163.
- Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11*, 315-319.
- Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin, 15*, 543-558.
- Foschi, M. (1989). Status characteristics, standards, and attributions. In J. Berger, M. Zelditch Jr., & B. Anderson (Eds.), *Sociological theories in progress: New formulations* (pp. 58-72). Newbury Park, CA: Sage.

- Foschi, M. (1992). Gender and double standards for competence. In C. L. Ridgeway (Ed.), *Gender, interaction, and inequality* (pp. 181-207). New York, NY: Springer-Verlag.
- Foschi, M. (2000). Double standards for competence: Theory and research. *Annual Review of Sociology*, 26, 21-42.
- Fuegen, K., & Biernat, M. (2010). *Inferring incompetence: How stereotypes affect decisions regarding inability*. Manuscript submitted for publication.
- Fuegen, K., & Endicott, N. F. (2009). *Evidence of shifting standards in judgments of male and female parents' job-related ability*. Manuscript submitted for publication.
- Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61-89). San Diego, CA: Academic Press.
- Hass, R. G., Katz, I., Rizzo, N., & Bailey, J. (1991). Cross-racial appraisal as related to attitude ambivalence and cognitive complexity. *Personality and Social Psychology Bulletin*, 17, 83-92.
- Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues*, 57, 657-674.
- Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin*, 28, 460-471.
- Levin, I. P., Rouwenhorst, R. M., & Trisko, H. M. (2005). Separating gender biases in screening and selecting candidates for hiring and firing. *Social Behavior and Personality*, 33, 793-804.
- Linville, P. W., & Jones, E. E. (1980). Polarized appraisals of out-group members. *Journal of Personality and Social Psychology*, 38, 689-704.
- Piacente, B. S., Penner, L. A., Hawkins, H. L., & Cohen, S. L. (1974). Evaluation of the performance of experimenters as a function of their sex and competence. *Journal of Applied Social Psychology*, 4, 321-329.
- Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, 74, 629-645.
- Rudman, L. A., & Fairchild, K. (2004). Reactions to counterstereotypic behavior: The role of backlash in cultural stereotype maintenance. *Journal of Personality and Social Psychology*, 87, 157-176.
- Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, 57, 743-762.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428.