Supporting Information

Moss-Racusin et al. 10.1073/pnas.1211286109

SI Materials and Methods

Subjects and Recruitment Strategy. To identify and screen potential participants, we used established practices similar to those used in other field experiments relying on nonundergraduate samples (1). We sought to strategically select departments for inclusion that were representative of high-quality United States science programs. Thus, participants were recruited from six anonymous American universities, all of which were ranked by the Carnegie Foundation as "large, Research University (very high research productivity)" (2). Additionally, each university had prominent, well-respected science departments (both at the undergraduate and graduate level), and tended to graduate high numbers of students who go on to careers in academic science. The schools were matched for size and prestige, and were selected from three different geographic regions within the United States. Within each region we included one private and one public university.

Within each university, participants were recruited from Biology, Chemistry, and Physics departments. These three fields were chosen because of their size, prominence, competitiveness, emphasis on research, and varying gender disparities. That is, all although all three showed gender disparities at the faculty level, the size of the gap differed. This gap was even more pronounced at the doctoral level, with some subfields of the biological sciences granting more doctorates to women than men (3). This diversity allowed for an examination of how faculty bias might differ as a function of the size and severity of the gender disparity across science fields. Although each institution had only one clear Chemistry and Physics department, some institutions had more than one Biology department (e.g., Ecology and Evolutionary Biology; Molecular, Cellular and Developmental Biology, and so forth). For such institutions, each of the core Biology departments were included. This method yielded a total of 23 departments.

For each selected department, departmental Web sites and publicly available course listings were used to create a full list of all eligible faculty participants. Potential participants had to meet several a priori qualifications to be eligible for participation. First, the participants had to be clearly identified as tenure-track faculty. As a result, Visiting Assistant Professors, Adjunct Professors, Instructional Staff, Research Faculty, Postdoctoral Associates, and Lecturers were not included. Additionally, participants had to be current faculty, thus excluding those with Emeritus titles. Faculty with primary appointments in other departments were also excluded, as were those whose appointments had yet to officially begin and those with invalid e-mail addresses. Finally, faculty who were identified as close personal friends or colleagues of one of the present study's authors were eliminated to avoid conflicts of interest. This method yielded a total of 547 eligible participants.

Data were collected during October and November of 2011. We followed the general methodological approach used in correspondence test audit research, typically used in field studies of discrimination (4, 5). All eligible participants received an e-mailed participation invitation originating from C.A.M.-R., indicated their consent, and completed all measures online. This method yielded a total of 165 participants, for an overall response rate of 30% (percentage rounded up). This percentage is on par with both similar empirical studies of professionals (6, 7) and that typically obtained in survey research (8). Additionally, extensive previous research has indicated that both demographic characteristics and substantive responses to focal variables largely do not differ between respondents and nonrespondents when sample demographics correspond to those of the underlying population (8). Thus, because the demographic information of our participants reflected the underlying population (as discussed below), the response rate obtained in the present study should allow for reasonably generalizable conclusions.

Data obtained from 30 participants were used to pilot and improve the study instruments and were thus not included in final analyses. Of the remaining 135 participants, 8 did not complete the majority of the study because of computer error (in three cases) or attrition (five cases); this resulted in a final sample of 127 participants for all substantive analyses. A power analysis indicated that this sample size exceeded the recommended $n = 90$ required to detect moderate effect sizes. Of participants, 74% were male and 81% were White (specific ethnic backgrounds were reported as follows: 81% White, 6% East-Asian, 4% South-Asian, 2% Hispanic, 2% African-American, 2% multiracial, and 1% each for Southeast-Asian, Middle-Eastern, and other), with a mean age of 50.34 (SD = 12.60, range 29–78). Of importance, these demographics are representative of both the averages for the 23 sampled departments (demographic characteristics for the sampled departments were 78% male and 81% White, corresponding closely with the demographics of those who elected to participate), as well as national averages (9). Additionally, 18% of participants were Assistant Professors, 22% were Associate Professors, and 60% were full professors, with 40% Biologists, 32% Physicists, and 28% Chemists. No demographic variables were associated with participants' substantive responses (all $P >$ 0.53). As expected when using random assignment, participants' demographic conditions did not vary across experimental conditions. Because there were 15 female and 48 male participants in the male student condition, and 18 female and 45 male participants in the female student condition, we obtained sufficient power to test our hypotheses (10).

Student Laboratory Manager Application Materials. We asked participants to rate a student laboratory manager application to help us develop appropriate mentoring programs for undergraduate science students. We prefaced the student laboratory manager application with text designed to bolster the credibility of the cover story and adjust for any differences in expectations and practices regarding laboratory managers between science fields (Fig. S1). Following conventions established in previous experimental work (11, 12), the laboratory manager application was designed to reflect slightly ambiguous competence, allowing for variability in participant responses and the utilization of biased evaluation strategies (if they exist). That is, if the applicant had been described as irrefutably excellent, most participants would likely rank him or her highly, obscuring the variability in responses to most students for whom undeniable competence is frequently not evident. Even if gender-biased judgments do typically exist when faculty evaluate most undergraduates, an extraordinary applicant may avoid such biases by virtue of their record. This approach also maintained the ecological validity and generalizability of results to actual undergraduate students of mixed ability levels.

Thus, we followed procedures established in previous similar research (11, 12) by designing an applicant who was "in the ballpark" for a laboratory manager position, but was not an obvious star. For example, although the applicant had completed 2 y of research experience and coauthored a journal article, their grade point average was slightly low (3.2) and they were described as having withdrawn from one class before the final. Fig. S2 displays the full text of the student laboratory manager

application materials for the female student condition. The sole difference in the male student condition was that the student's name read as "John" instead of "Jennifer," and female pronouns were replaced with male pronouns.

To ensure that the application materials reflected the desired degree of competence, they were developed in consultation with a panel of academic science researchers who had extensive experience hiring and supervising student research assistants. After the materials were developed, they were then rated by a separate group of knowledgeable graduate students, postdoctoral scholars, and faculty. Results from this pilot testing revealed consensus that, as intended, the materials reflected a qualified but not irrefutably excellent applicant.

Dependent Variable Scales. Participants completed the following scales, which were well-validated and modified for use from previous studies (13–15).

Student competence. The target student's competence was assessed using three items on a 1 (not at all) to 7 (very much) scale. These items were: (*i*) Did the applicant strike you as competent? (*ii*) How likely is it that the applicant has the necessary skills for this job? (*iii*) How qualified do you think the applicant is? ($\alpha = 0.93$). Items were averaged to form the student competence scale, with higher numbers indicating greater levels of perceived competence.

Student hireability. The extent to which the student applicant was viewed as hireable for a laboratory manager position was measured using three items on a 1 (not at all likely) to 7 (very likely) scale. These items were: (*i*) How likely would you be to invite the applicant to interview for the laboratory manager job? (*ii*) How likely would you be to hire the applicant for the laboratory manager job? (*iii*) How likely do you think it is that the applicant was actually hired for the laboratory manager job he/she applied for? ($\alpha = 0.91$). Items were averaged to compute the student hireability scale, such that higher numbers reflected greater perceived hireability.

Salary conferral. Salary conferral was measured using one item, If you had to choose one of the following starting salaries for the applicant, what would it be? Responses were indicated on the following scale: 1 ($15,000), 2 ($20,000), 3 ($25,000), 4 ($30,000), 5 ($35,000), 6 ($40,000), 7 ($45,000), 8 ($50,000). Collapsed across conditions, the average recommended salary was $28,373.02 (SD = $6,382.14), with a range of $15,000 to $45,000.

Mentoring. The extent to which participants were willing to mentor the student applicant was assessed using three items on a 1 (not at all likely) to 7 (very likely) scale. These items were: If you encountered this student at your own institution, how likely would you be to... (*i*) Encourage the applicant to stay in the field if he/she was considering changing majors? (*ii*) Encourage the applicant to continue to focus on research if he/she was considering switching focus to teaching? (*iii*) Give the applicant extra help if he/she was having trouble mastering a difficult concept? ($\alpha = 0.73$). Items were averaged to form the mentoring scale, with high numbers reflecting greater willingness to mentor the student.

Subtle gender bias. Our intention was to select a scale that would measure modern bias against women. We reasoned that, as mentors and educators, many faculty members would not report high levels of "old-fashioned" or hostile sexism, characterized by overtly negative evaluations of women and the desire to halt women's progress (16). Instead, we were interested in how shared subtle gender biases (resulting from pervasive cultural messages) might impact perceptions of female students. As a result, we sought a measure that would tap a subtle, modern form of gender bias that often exists outside of individuals' conscious awareness or intention to harm women. Thus, we used the Modern Sexism Scale, a commonly used and well-validated scale that functions as an indirect measure of modern views toward women and gender (17). Participants responded to eight items on a scale ranging from 1 (strongly disagree) to 7 (strongly

agree). Items included: On average, people in our society treat husbands and wives equally; Discrimination against women is no longer a problem in the United States; and Over the past few years, the government and new media have been showing more concern about the treatment of women than is warranted by women's actual experiences ($\alpha = 0.92$). Items were averaged to form the gender attitudes scale, with higher numbers indicating more negative attitudes toward women.

Likeability. Using a scale ranging from 1 (not at all likely) to 7 (very likely) scale, participants answered three items indicating the extent to which they liked the student applicant. These were: (*i*) How much did you like the applicant? (*ii*) Would you characterize the applicant as someone you want to get to know better? (*iii*) Would the applicant fit in well with other laboratory members? Items were averaged to create the likeability scale, with higher numbers representing greater liking of the target student ($\alpha = 0.87$).

Analytic Strategy. Although we treated the individual as the primary unit of analysis for all statistical tests, it should be noted that our data may be conceptualized as clustered or nested by various groups (18). That is, because participants belonged to one of three distinct Science fields (Biology, Chemistry, or Physics) and also one of 23 distinct departments, their membership in a science field or a department could result in nonindependence of observations on at least one of these levels. For example, if bias against female undergraduates is systematically greater among chemists than biologists, then our data would be nonindependent at the level of science field (in that scores on any given dependent variable would be more similar for two chemists than two randomly selected individuals) (18). Because standard inferential statistics assume that all observations are independent, we may have introduced error by failing to account for the nested nature of our data (i.e., nonindependence because of groups) (19).

To address this issue, we followed recommended practices to assess the possible nonindependence of our data (20). If data are found to be independent at the level of a nesting variable, then it is acceptable to not account for this variable in statistical tests (18, 19). To evaluate the nonindependence of our data, we conducted a null multilevel model for each dependent variable. Using the MIXED procedure in SPSS, we included both predictors (student and faculty gender, effects-coded such that male = −1, female = 1) in each model. For each model, the intraclass correlation coefficient was near zero and nonsignificant (all $P > 0.11$), suggesting that our data did not violate assumptions of independence. Put another way, there was no significant variance associated with participants' group membership in a given science field. As a result, we concluded that it was appropriate to analyze our data without accounting for the impact of this nesting variable.

One additional potential source of group level variance remains. Our data were also nested at the level of department. Although it is possible that participants' scores were nonindependent based on departmental membership, we were not able to estimate nonindependence at this level because of ethical concerns that shaped our recruitment and data-collection strategies. That is, following the stipulation of the reviewing Institutional Review Board, we assured full anonymity to our faculty participants. As a result, it was impossible to ask them to indicate to which department they belonged. We agreed with the Institutional Review Board's assessment that participation and attrition rates as well as responses to potentially sensitive questions may have been biased if participants feared that their identities could be gleaned from the information they provided. Indeed, among (sometimes small) academic fields, knowing an individual's university, department, and demographic characteristics would likely be sufficient information to identify them personally.

Thus, to avoid undermining recruitment strategies and biasing participants' responses, we refrained from collecting information

about specific departmental membership. As such, we were unable to determine whether responses were nonindependent as a function of this variable, and if so, to account for it in our statistical approach. Ignoring significant nonindependence does not bias effect estimates, but can bias degrees of freedom, variances, SEs, and significance tests (20). Monte Carlo simulations have determined that under these circumstances, the resulting SEs may be too large, too small, or hardly biased, resulting in possible type 1, type II, or no errors (21). However, because the data were shown to be nonindependent at the level of science field and there is no a priori theoretical reason to predict nonindependence at the level of department, we proceeded with a standard inferential statistical approach. Nonetheless, future research should seek to develop methods to measure and control for potential nonindependence due to departmental membership.

Additional Analyses. *Mediation analyses.* To test for mediation, we followed procedures recommended by Baron and Kenny (22). Student gender was coded 0 (male), 1 (female). As noted in the main text, primary mediation analyses evaluating hypothesis C were conducted using the composite competence variable, following established best practices (e.g., refs. 23–26). However, we also ensured that results were similar using the noncomposite competence scale (i.e., without the salary conferral variable), to rule out the possibility that the mediation results were driven solely or primarily by the salary conferral variable. As expected, results using the noncomposite competence scale were similar although slightly weaker, in that the initially significant relationship between student gender and student hireability ($\beta = -0.35$, $P < 0.001$) was reduced in magnitude and significance ($\beta = -0.13$, $P = 0.05$) after accounting for the impact of student competence (which was a strong predictor, $\beta = 0.69$, $P < 0.001$), Sobel's $Z = 3.65$, $P < 0.001$. These results suggest that the composite competence variable functioned as expected.

Additionally, although not specifically predicted, we examined whether the composite competence variable might also mediate the relationship between student gender and mentoring. Results demonstrated partial mediation, in that the initially significant relationship between student gender and mentoring ($\beta = -0.32$, $P < 0.001$) was reduced in magnitude and significance ($\beta = -0.22$, $P = 0.02$) after accounting for the impact of student composite competence (which was a significant predictor, $\beta = 0.28$, $P < 0.05$), Sobel's $Z = 2.91$, $P < 0.01$.

Because mentoring has strong social components and may be perceived as less immediately task-relevant than hireability, we did not initially expect it to be mediated by the composite competence scale, and this could account for why we observed partial rather than full mediation with this variable (relative to the full mediation observed for hireability, supporting hypothesis C). However, the fact that evidence for partial mediation emerged even for mentoring (a secondary downstream dependent variable in the current context) speaks to the powerful impact of differences in the perceived competence of male and female students.

Moderation analyses. To test for moderation, we first standardized all variables and then ran a series of multiple regression analyses with student gender, faculty participants' negative attitudes toward women, and their interaction predicting student composite competence, hireability, and mentoring. As noted, the interaction was a significant predictor in each case. As with mediation, we next ensured that results were similar using the noncomposite measure of competence (to determine that moderation results for this variable were not driven solely by the salary conferral variable, which was included in the composite competence measure). Results of multiple regression analyses indicated that participants' preexisting subtle bias against women significantly interacted with student gender condition to predict perceptions of student noncomposite competence ($\beta = -0.42$, $P < 0.01$). As expected, bivariate analyses results revealed that the more preexisting subtle bias participants exhibited against women, the less noncomposite competence ($\beta = -0.38$, $P < 0.01$) they perceived the female student to possess. In contrast, faculty participants' levels of preexisting subtle bias against women were unrelated to the perceptions of the male student's noncomposite competence ($\beta = 0.18$, $P = 0.16$). These results are nearly identical to those obtained with the composite competence index, suggesting that those findings were not driven solely by the salary conferral variable and providing additional evidence that the composite competence variable functioned as intended.

We then explored whether additional variables might interact with participants' subtle preexisting bias against women to predict their reactions to the target students. Separate models adding faculty participant gender, age, science field, tenure status, each two-way interaction, as well as the three-way interaction of each demographic variable with student gender condition and faculty participant gender (to rule out participant gender differences) revealed no significant novel predictors (all $\beta < 0.38$, all $P > 0.28$). This finding suggests that faculty participants' gender attitudes themselves played a role in undermining support for the female (but not male) student, and that the impact of these gender attitudes does not appear to vary as a function of participants' other demographic characteristics, including their gender. Consistent with other results, it appears that female as well as male faculty members' negative attitudes toward women undermined their support for the female student, irrespective of their age, science field, and career status.

1. Feldon DF, et al. (2011) Graduate students' teaching experiences improve their methodological research skills. *Science* 333:1037–1039.
2. Carnegie Foundation for the Advancement of Teaching (2012) *Standard listing of research universities (very high research productivity)*. Available at http://classifications. carnegiefoundation.org/lookup_listings/srp.php?clq=%7B%22basic2005_ids%22%3A% 2215%22%7D&start_page=standard.php&backurl=standard.php&limit=0,50. (Accessed February 14, 2012).
3. National Science Foundation (2009) *Women, Minorities, and Persons with Disabilities in Science and Engineering* (National Science Foundation, Arlington).
4. Pager D (2007) The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *Ann Am Acad Pol Soc Sci* 609 (1):104–133.
5. Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am Econ Rev* 94: 991–1013.
6. Steinpreis RE, Anders KA, Ritzke D (1999) The impact of gender on the review of curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles* 41:509–528.
7. Wunder GC, Wynn GW (1988) The effects of address personalization on mailed questionnaires response rate, time, and quality. *J Mark Res Soc* 30(1):9–101.
8. Holbrook AL, Krosnick JA, Pfent A (2007) *Advances in Telephone Survey Methodology*, eds Lepkowski JM, et al. (John Wiley & Sons, Hoboken), pp 499–528.

9. National Science Foundation (2008) *Survey of Earned Doctorates* (National Science Foundation, Arlington).
10. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22:1359–1366.
11. Foschi M (2000) Double standards for competence: Theory and research. *Annu Rev Sociol* 26(1):21–42.
12. Heilman ME, Wallen AS, Fuchs D, Tamkins MM (2004) Penalties for success: Reactions to women who succeed at male gender-typed tasks. *J Appl Psychol* 89:416–427.
13. Moss-Racusin CA, Phelan JE, Rudman LA (2010) When men break the gender rules: Status incongruity and backlash against modest men. *Psychol Men Masc* 11(2):140–151.
14. Moss-Racusin CA, Rudman LA (2010) Disruptions in women's self-promotion: The backlash avoidance model. *Psychol Women Q* 34:186–202.
15. Rudman LA, Moss-Racusin CA, Phelan JE, Nauts S (2012) Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice toward female leaders. *J Exp Soc Psychol* 48(1):165–179.
16. Glick P, Fiske ST (1996) The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *J Pers Soc Psychol* 70:491–512.
17. Swim JK, Aikin KJ, Hall WS, Hunter BA (1995) Sexism and racism: Old-fashioned and modern prejudices. *J Pers Soc Psychol* 68:199–214.
18. Kenny DA (1996) The design and analysis of social-interaction research. *Annu Rev Psychol* 47:59–86.

19. Kenny DA, Judd CM (1986) Consequences of violating the independence assumption in analysis of variance. *Psychol Bull* 99:422–431.
20. Kenny DA, Judd CM (1996) A general procedure for the estimation of interdependence. *Psychol Bull* 119:138–148.
21. Hox J (1998) *Classification, Data Analysis, and Data Highways*, eds Balderjahn I, Mather R, Schader M (Springer, New York), pp 147–154.
22. Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 51:1173–1182.
23. Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 56:81–105.
24. Robins RW, Hendin HM, Trzesniewski KH (2001) Measuring global self esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Pers Soc Psychol Bull* 27(2):151–161.
25. Moss-Racusin CA, Rudman LA (2010) Disruptions in women's self-promotion: The backlash avoidance model. *Psychol Women Q* 34:186–202.
26. Rudman LA, Moss-Racusin CA, Glick P, Phelan JE (2012) *Advances in Experimental Social Psychology*, eds Devine P, Plant A (Elsevier, New York), pp 167–227.

In the current phase of the study, we are interested in how undergraduate students' are selected for competitive lab manager positions following their graduation and before they apply to doctoral programs. Lab managers are an important part of the scientific enterprise, but very little is know about how people weigh various factors and make the complicated decision to hire one lab manager from a pool of applicants.

*To study this question, we have compiled and summarized information from **actual applications of students who have recently applied to be lab managers** at universities across the country. These students have volunteered to share their information in exchange for mentoring opportunities as part of their participation in the study. We have summarized their information by creating profiles for each applicant using a standardized form, in order to adjust for individual differences in application procedures and enable consistent evaluations across applicants.*

*Today, we will be assigning you to read the applicant profile of one randomly-selected student from the nationwide database. Please imagine that you are actually evaluating this student's application to work **in your own lab**. After reading the applicant profile, you will be asked to provide your opinions of the student, and offer them feedback as they make decisions about moving forward with their career. These students come from fields where lab managers are typically students with research-focused career ambitions who work in these types of positions for several years before going on to graduate school. Even if your particular sub-field does not have such practices regarding lab managers, please evaluate the student within this context.*

Remember, the national database we have created contains information both from very qualified students, as well as students who are less competitive. Thus, we ask that you be prepared to offer honest feedback to the student you are selected to read about, regardless of how qualified they are. Please do not hold back, and don't be afraid to be truthful—the success of the current project depends upon your ability to provide honest, straightforward feedback.

Fig. S1. Cover story text. The text in the figure was viewed by participants in PDF format without additional supporting text.

DEMOGRAPHICS
Participant ID #: 149
Name: Jennifer ▮▮▮▮
Gender: Female
Ethnic Background: Caucasian
Age: 22
Degree: Bachelors of Science, obtained May 2011 from ▮▮▮▮▮▮ University

BACKGROUND
GPA: 3.2
GRE score: 650 verbal, 780 quant
Awards/honors: President's Service Award, Rotary Club College Scholarship
Previous research experience: 2 years as a research assistant working with 2 different faculty mentors
Academic standing: appears from Jennifer's transcript that she was in good standing upon graduation, but withdrew from 1 class prior to final
Letters of recommendation: 3 (2 from former faculty research supervisors, 1 from an intro science course professor), all supportive
Future plans: apply to doctoral programs
Extracurricular activities: student government, college learning center tutor
Position sought: Lab Manager
Position duration: 2 years, with possibility of renewal pending satisfactory performance

STATEMENTS/LETTERS
Excerpt from student statement: "I am a motivated student and would make the most of the opportunity to serve as your lab manager. After spending a semester working in Dr. ▮▮▮▮'s lab and another year doing research with Dr. ▮▮▮, I have gained valuable technical skills, co-authored a journal article, and am now committed to an academic research career…as someone focused on improving my standing and enhancing my research experience, this lab manager position would provide the perfect opportunity to hone the necessary skills to make me competitive for graduate school applications… additionally, the fascinating research taking place in your lab is directly in line with my interests and experiences…in short, I am focused, motivated, organized and dedicated to improving my research skills. I am enthusiastic about the opportunity to fill the lab manager position and collaborate with you on future research."

Excerpt from faculty recommendation letter: "…although Jennifer admittedly took a bit longer than some students to get serious about her studies early in college, she has impressed me by improving over the last two years of her science coursework and has made every effort to make up for lost ground…she has been a strong research assistant in my lab, and I know she is capable of serving as a dedicated lab manager."

Fig. S2. Lab manager application materials (female student condition). The only differences in the male student condition were that the name "Jennifer" was replaced with "John," and all female pronouns were replaced with male pronouns.