PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES
SUPPORTING INFORMATION APPENDIX FOR:
WILLIAMS, W.M. & CECI, S.J. (2015). NATIONAL HIRING EXPERIMENTS REVEAL 2:1 FACULTY PREFERENCE FOR WOMEN ON STEM TENURE TRACK

SEE VIDEOS AND ADDITIONAL RESOURCES AT WWW.CIWS.CORNELL.EDU

TABLE OF CONTENTS

PREFACE

This Technical Supplement contains a comprehensive description of this program of research. It begins with a justification and explication of the research question. Next, the supplement describes various validity checks undertaken to support the representativeness of the sample and the generalizability of the findings. This portion includes a discussion of the sampling plan, why we oversampled Ph.D.-granting institutions, and why the denominators used in our calculation of sample weights exclude institutions that were inappropriate for inclusion in the experiment (either because they lacked tenure-track faculty in the fields of interest or because they were not a part of higher education--e.g., stand-alone culinary institutes, seminaries, Vo-Tech programs). Following this discussion, we describe the research design, explaining each cell in it and describing the task as it was presented to faculty respondents. Finally, we present the statistical findings in detail--including describing a full reanalysis of the unweighted data using sample weights, which confirmed the results from the unweighted analysis reported in the main article. Note that due to the extensive nature of this program of research (five experiments, spread over four years) it was not possible to provide descriptions of every aspect of every experiment in this Supplement, due to PNAS's 30-page limit. Therefore, we have archived "Supporting Information Resources" and additional data files at the Cornell Institute for Women in Science webpage. Please see: http://www.ciws.cornell.edu.

I. Research Question, Experimental Design, Sampling Plan, and Method

a. Gender Bias in Tenure-Track-Hiring Preferences in Math-Intensive Fields. There have been hundreds of studies of sex bias in hiring and work-product evaluation, including an early meta-analysis of 49 such studies (1) and many dozens of studies since then, including a number of experimental studies. However, the present study is the only experimental study of sex bias in tenure-track hiring in math-intensive fields, and it is the only study to include a large, nationally representative sample of tenure-track faculty from doctoral, master's, and baccalaureate institutions, as explained below. No prior study has experimentally examined tenure-track-hiring bias in math-intensive fields and employed a large, national sample of tenure-track faculty from various types of institutions, with multiple validity checks to ensure its representativeness and generalizability and to rule out plausible alternative hypotheses. The present experiment was designed to shed light on whether male and female Ph.D. applicants with identical scholarly records (differing solely in the gender pronoun used to describe them) were viewed as equally hirable by tenure-track faculty across the U.S., at small teaching-intensive colleges to large research-intensive universities, and whether identical male and female applicants were viewed as equally hirable in both math-intensive fields (engineering, economics) and non-math-intensive fields (biology, psychology), and in fields in which women are well-represented (biology, psychology) and those in which they are not (engineering, economics).

The reason this latter question is important is because it is in the math-intensive fields that women are most underrepresented (2), yet hiring bias has never been studied experimentally in these fields. In non-math-intensive fields, women now represent substantial portions of faculty. However, despite making progress in mathematical fields, women continue to be highly underrepresented as a fraction of total faculty. Is this due to discriminatory hiring preferences, a shortage of women earning degrees and applying for tenure-track positions, or some other reason? To address this question, four fields were chosen for study to determine whether the ones that are math-intensive and in which women are a small fraction of the tenure-track faculty (economics and engineering) differ in their hiring preferences from the ones in which women are well-represented (psychology and biology). Generally, women are well represented outside the math-intensive domain, including in psychology where women comprise nearly half of assistant professors (and 38% of professors across all ranks) and over a third of those hired in biology. Based on the 2010 NSF Webcaspar data, women comprise 48.1% of all tenure-track assistant professors in biological and social science fields, but only 29.6% of assistant professors in math-intensive fields (2, Figures 1b and 3a).

Although a great deal has been written about women's difficulties getting tenure-track appointments in math-intensive fields, none of this writing has been grounded in controlled experiments contrasting identical applicants in math-based fields. This lack of data represents an unfortunate missing piece of the puzzle. Conclusions based on non-experimental research in math-intensive fields are open to alternative explanations due

to potential unmeasured variables and the lack of control for the most important variable--*applicant quality*--which is, at best, proxied in some writings by total number of publications without regard to their quality. Ideally, conclusions should derive from a test of the hirability of two otherwise identical applicants, one of each gender. Using a "between-subjects" experimental design with random assignment, causal inferences can be made about the reasons for women's underrepresentation in math-intensive tenure-track positions. Will tenure-track faculty members prefer male applicants over identical female applicants, as the non-experimental literature often claims? Of course, experiments using random assignment carry their own causal limitations, most notably validity and generalization threats. As described below, we addressed such concerns in several ways, including using mixed methods that combine experimental data with correlational data, recruiting a paid validation sample that approached 100% participation, and by using sample weights to adjust for nonrandomness. Collectively, these procedures give us a reasonable degree of confidence that the findings reported here are valid and generalizable. Of course, no findings are perfect and it is always possible to find features of an experimental study that could be improved. This study is no exception (e.g., we did not examine all STEM fields and we studied only hiring at the entry assistant professor stage—although elsewhere we have done extensive analyses of post-hiring sex differences in salary, promotion, job satisfaction, persistence, impact, and grant awards; see 2). However, this study has important advantages over others that have been discussed in policy circles, representing superior design, sampling, and follow-up experimentation.

b. Experimental Design. Our experimental design incorporated three primary factors—gender of hypothetical job applicants, gender of actual tenure-track faculty who ranked these hypothetical applicants, and whether faculty are from math-intensive or non-math-intensive fields. In addition, various lifestyle variations of applicants (married, divorced, with or without young children, etc.) that could influence faculty-hiring preferences were contrasted in 12 conditions across five experiments. In Experiment 1, six conditions with two subconditions each (1.1/1.2-6.1/6.2) crossed the three primary variables mentioned above in a fully counterbalanced design involving 363 faculty who rated male and female applicants with the same lifestyle (e.g., single, no children). In Experiment 2, two additional conditions (7.1/7.2 and 8.1/8.2) tested 144 faculty raters' preference for hypothetical applicants who differed in their lifestyles (e.g., single, divorced mother competing against traditional father with stay-at-home wife). In every case in these two experiments (conditions 1.1/1.2 through 8.1/8.2, N=507 faculty), the contest presented to every faculty member involved a man and a woman with comparable qualifications, plus one additional man, the foil, who had slightly-weaker qualifications, all competing for the same tenure-track assistant professorship. The foil candidate received only 2.53% of first-place rankings across our entire study, being viewed as we had pretested him to be, but still representing a realistic applicant to be on a short list. (Below we describe how the foil was pretested.) These contests between equivalently-qualified male and female applicants were further disguised by describing applicants' personae differently, using adjectives shown in past research to be associated with femaleness as opposed to maleness: kind, creative, socially-skilled vs. independent, analytic, powerhouses. These personae were fully counterbalanced, so that half of faculty received a male candidate depicted with one persona and the female depicted with the other, and the other half of faculty received these personae with gender of candidate swapped. Thus, the use of these personae disguised the main hypothesis by leading faculty to believe the research question involved their relative preference for these two types of personae (see below), but they were counterbalanced to control for any differences in faculty's preferences. In Experiment 3, two conditions (9.1/9.2 and 10.1/10.2) examined the effect of taking a one-year parental leave while in graduate school. We did this by contrasting two <u>same-sex</u> applicants with identical records, one of whom took a leave—and we used an opposite-gender, slightly-weaker foil candidate to round out a realistic slate of three shortlisted candidates.

To recap, in Experiment 1, which included the first 6 experimental conditions, the competition was between a man and woman with the same lifestyle and a slightly weaker foil candidate, rated by faculty who themselves were half women and half men (Conditions 1.1/1.2-6.1/6.2). Experiment 2 involved four conditions—7.1/7.2 and 8.1/8.2—in which the competition pitted a married father of two with stay-at-home spouse against a divorced mother of two (7.1/7.2) or against a single woman with no children (8.1/8.2), both contrasts relevant to real-world situations and formerly suggested in the literature as eliciting bias against females (i.e., a pro-married-father-with-stay-at-home-wife effect). Experiment 3 included four experimental conditions in which the competition was between either two women (and a male foil), one of whom took a leave (Conditions 9.1/9.2) or between two men (and a female foil), one of whom took a leave (Conditions 10.1/10.2).

In Experiment 4 a small group of 35 engineering professors were asked to do the same task their peers did in the earlier experiments but instead of giving them narrative summaries of the applicants they were given full CVs. This was done to see whether the same results would be found. As was the case with their colleagues who were given narrative summaries, faculty given full CVs preferred the female applicant over the identically-qualified male, 75.8% vs. 24.2%.

In Experiment 5 we examined whether the same results would be found if instead of asking faculty to choose between a male and female applicant for the position they were asked to evaluate just one applicant, either a female or a male, in the absence of a contest between opposite sex applicants which might induce politically-correct responding to prefer the female even if the faculty held implicit biases against her. Would their implicit biases come to the fore and lead them to downgrade the female vis-à-vis ratings given by their colleagues to male applicants, as some have claimed? In this experiment, 127 faculty from the fields of biology and engineering were asked to evaluate only one candidate, either a male or female (who were identical except for gender pronoun), using a 10-point rating scale. Females were rated significantly higher than males, 8.20 vs. 7.14, $F(1,123)=16.48$; $p < .001$, suggesting, as we noted in the text, that values favoring gender diversity have become internalized among U.S. faculty in these fields.

c. Disguising the Research Hypotheses: Use of Adjectives to Create Gendered Personae. The design of this experiment fully crossed "gendered personae", so that half of the Dr. Xs and Zs were depicted in stereotypically male terms (e.g., powerhouse, analytic, independent, competitive, stands up under pressure, single-minded) and half were depicted in stereotypically female terms (e.g., creative, original, imaginative, likeable, kind, socially-skilled). In other words, half of male applicants were described in a stereotypically male persona and the other half were described in a stereotypically female persona (3, 4), and the same was true of female applicants, with half described in each gendered persona. Since these personae were counterbalanced across faculty raters, they did not influence the overall results.

The systematic manipulation of gendered personae served a very important role of subterfuge in the present experiments. Because one applicant was depicted in a female gendered persona and the other--although identical in academic accomplishments--was depicted in a male persona, it suggested to respondents that we were interested in their relative preference for creative, imaginative, kind, socially-skilled individuals vs. analytic, single-minded, ambitious, powerhouses, among otherwise identically-qualified applicants (identically rated as 9.5 with strong letters, eminent mentors, and strong department evaluations). Gendered personae were fully crossed with gender of applicant (e.g., each male applicant presented to a faculty rater in a stereotypically male persona was also presented to a different faculty rater in a female persona, counterbalanced for faculty-rater gender as well). Thus, respondents were unaware of this controlled manipulation (since it was done between subjects), and if they harbored any hunches about the purpose of the experiment, they reported that they assumed our interest was in their preference between these types of personae, which served to disguise the gendered nature of the experiment. Coupled with the inclusion of a male (and sometimes female) Y foil, this manipulation of personae resulted in faculty believing the purpose of the experiments was to determine their preference for creative, kind, socially-skilled applicants vs. analytic, single-minded, ambitious powerhouses--not to test their preference for males versus females. In the next section (**IIe**) we report a validation of the gendered personae.

d. Large National Sample. The pool of potential faculty for Experiments 1, 2 and 3 was assembled by drawing a national stratified sample of 2,090 professors (half female, half male, across all ranks). This was done by randomly sampling from online directories for Carnegie Foundation's 3 Basic Classifications of: a) Doctoral (combining all three levels of doctoral intensity), b) Master's/Baccalaureate Combined institutions (combining all three levels—small, medium, and large), and c) Baccalaureate institutions (combining all three levels of such institutions). This sample of 2,090 professors was drawn equally from four popular fields, two math-intensive in which women faculty are greatly underrepresented-- $\leq$15% (engineering, economics)--and two non-math-intensive (biology, psychology), in which women faculty are well represented and are considered to have achieved what gender equity advocates regard as a critical mass, although even these fields still produce significantly more female PhDs than the female fraction of total professorships. A constraint in randomization was that for an institution to be included it had to have programs in at least three of the four fields. This was true of all doctoral institutions in the sampling frame but it excluded many small colleges that lacked

two or more of the four fields and over half of the nation's combined master's programs as we describe in the Resources section of this Supplement at http://www.ciws.cornell.edu. Institutions were selected randomly with the stipulation that PhD-granting institutions represent half the sample. Oversampling of these institutions was desired because of their prestige, significantly higher average salaries (e.g., AAUP, March/April, 2013, Survey Report, http://www.aaup.org/sites/default/files/files/2014%20salary%20report/Table4.pdf), larger faculties and student bodies, and because they historically have been bastions of female underrepresentation in math-intensive fields. Within each of the sampled institutions, one male and one female tenured or tenure-track faculty member was randomly selected in each of the four fields, plus additional male and female faculty were randomly chosen as replacements in the event of nonresponse and additional faculty were selected for inclusion in follow-up validity studies. Out of 2,090 tenure-track faculty who were solicited in this manner, 711 provided data, yielding an overall response rate of 34.02%, and an additional 35 faculty provided validation data on two of the cells, as will be described in **IIc**. The response rates for the 3 types of institutions x 2 genders x 4 fields ranged between 24.5% and 43.2% (see Table S1 below). A weighted analysis that adjusted for the response rates did not alter any findings (see Resources section at http://www.ciws.cornell.edu). In Experiment 1 stratified random sampling within gender, type of institution (Carnegie 1, 2, 3), and discipline was used. However, in follow-up experiments there was no formal randomization within strata. Although Experiments 2, 3, and 5 constituted geographically broad samples, they were not formal national probability samples in the sense that political pollsters use this term. The goal was to probe causal mechanisms responsible for the effects observed in Experiment 1 rather than make claims of a national probability sample. Since the effects in the first experiment were quite large, we abandoned the time and expense of continued stratification in the next four experiments.

The total sample for the first three experiments was 711 faculty, plus an additional sample of 35 professors of engineering was collected for Experiment 4 to test the effects of giving faculty full CVs instead of narrative summaries. An analysis of these engineers' responses showed them to be indistinguishable from those of their colleagues in the main sample of 711, thus validating the procedure. For purposes of reporting here, all analyses are based on the 711 faculty from Experiments 1-3 (the main sample), all of whom followed the identical procedure. Statistical tests show that only one of the findings from this main sample of faculty respondents is changed by including the added sample of 35 engineering professors: error bars surrounding male engineering faculty's preference for female applicants over identically-qualified male applicants no longer (slightly) overlap when these extra faculty are included as seen in Fig. 1. Thus, with the sole exception of male economists, who had no preference between identically-qualified male and female applicants, all other groups of male and female faculty, across all fields, strongly preferred female applicants and the error bars reflecting their female preferences do not overlap with the bars representing their male preferences.

As mentioned, equal numbers of male and female faculty were solicited, so it is of interest that the response rates for male and female faculty were also virtually equal: Of the 711 respondents in the first three experiments who provided complete data, 355 were females and 356 were males. When Experiments 4 and 5 are added, the split is 434 female faculty and 439 males. Such mirroring of the 50-50 original sample was unexpected. Tenure-track faculty were confined to assistant, associate, and full professors, and chair-holders who were not further qualified by designations such as "visiting", "term", "courtesy", "emeritus", etc. Moreover, adjuncts, instructors, lecturers, and senior lecturers were not included in the sampling nor were they included in the calculation of sample weights described in Section **II**. This is because many people in these categories are not involved in hiring tenure-track faculty, which is the core of what respondents were charged with doing in our task. Data were collected beginning in the second half of 2011 for Conditions 1.1/2-6.1/2; data collection extended into 2012 for Conditions 7.1/2-10.1/2; full-CV engineering and psychology paid validation samples were run from the second half of 2012 into the first half of 2013; the final experiment was conducted in late Fall 2014.

Table S1: Response rates (percent of faculty solicited who provided data) for the three types of Carnegie institutions as a function of gender of respondent and field (department). Response rates do not include paid subsample of 82 psychologists with 91.1% response rate.

	Carnegie 1 (Doctoral)		Carnegie 2 (Master's/BA)		Carnegie 3 (Small Colleges)	
	Male	**female**	**male**	**female**	**male**	**female**
Biology	39.0%	28.8%	26.9%	34.5%	25.0%	30.7%
Economics	33.8%	24.0%	18.4%	26.5%	37.8%	43.2%
Engineering	28.1%	31.7%	23.5%	31.1%	35.3%	23.1%
Psychology	32.7%	35.6%	30.5%	33.3%	34.2%	35.7%

e. Methods. For Experiments 1-3, equal numbers of male and female faculty respondents were randomly assigned to one of 20 "between-subjects" conditions, with conditions counterbalanced so the gender of the hypothetical applicants in one version that was sent to some faculty was switched to the opposite gender in the counterbalanced version that was sent to other faculty. Below we provide an example of such a gender-switch in materials sent to different faculty respondents. Each faculty respondent was given descriptions of just three hypothetical applicants (denoted Drs. X, Y, and Z, rather than by names to avoid any unintended connotations that names might carry). Faculty were asked to rank Drs. X, Y, and Z for a tenure-track assistant professorship in their own department. Two of these three applicants in each of the 20 conditions (Drs. X and Z) were depicted as slightly stronger than the third applicant (Dr. Y), a foil. This was conveyed through information about the three applicants that each faculty respondent received: faculty respondents were informed that both Drs. X and Z had been rated by faculty in their department 9.5 on a scale in which 10 is outstanding/exceptional; they were informed that the third applicant, Dr. Y, had been rated slightly lower, but still strong (9.3) and Y was described slightly less enthusiastically in the search committee chair's notes, though still very positively. (See examples of materials sent to respondents in the Resources section at http://www.ciws.cornell.edu**.**)

The inclusion of a slightly weaker Dr. Y "foil" candidate was done to obscure the gendered nature of the study. That is, if faculty had been asked to rank a single male against a single female applicant, the gendered nature of the comparison might have been evident, particularly in math-intensive fields in which searches usually contain mostly male applicants; by including a Y foil who was male, this resulted in 2 males and one female, possibly rendering the gendered nature of the experiments less obvious. A second ploy to obscure the gendered nature of the experiments was that Drs. X and Z were described using opposite-gendered personae, which created realistic applicant-depictions. These gendered personae were swapped so that half of all female applicants were described with stereotypically-female gendered adjectives and half were described with stereotypically-male gendered adjectives, and the same was true of the male candidates—half were described with female adjectives and half with male adjectives. By asking faculty to rank three applicants, two-thirds of whom were male and all of whom differed in the adjectives used to describe them, it was not obvious to raters that this was a study of gender preference because they were unaware that other faculty were sent mirror reversals of the descriptions they were sent. And this ploy worked; when a subset of 30 respondents was asked to guess the hypothesis of the study, none suspected it was related to applicant gender.

As expected, the weaker male foil, Dr. Y, was ranked first (most preferred to be hired) by only 2.53% of faculty respondents, so his inclusion helped serve the purpose of a foil who obscured the gendered nature of the experiment while not diverting data from Drs. X and Z, the two real contestants whose credentials were identical (9.5) and whose adjectival personae were identical (i.e., reversed) between-subjects, differing only in their gender, conveyed simply by referring to them as she or he.

In the Resources section are samples of lifestyle conditions (each containing two subconditions) to which faculty were randomly assigned, each of which was counterbalanced to reverse the gender of Drs. X and Z. (http://www.ciws.cornell.edu.) As an example of how the information about the applicants, Drs. X, Y, and Z, was conveyed to faculty respondents, the following is a portion of the text given to faculty respondents in one of the twenty conditions, describing the applicant Dr. X, who in this version happens to be a female with

stereotypically-male persona (analytic, powerhouse), who is single without children:

Imagine you are on your department's personnel/search committee. Your department plans to hire one person at the entry assistant-professor level. Your committee has struggled to narrow the applicant pool to three short-listed candidates (below), each of whom works in a hot area with an eminent advisor. The search committee evaluated each candidate's research record, and the entire faculty rated each candidate's job talk and interview on a 1-to-10 scale; average ratings are reported below. Now you must rank the candidates in order of hiring preference. Please read the search committee chair's notes below and rate each candidate. The notes include comments made by some candidates regarding partner-hire and family issues, including the need for guaranteed slots at university daycare. If the candidate did not mention family issues, the chair did not discuss them.

Dr. X: *X struck the search committee as a real powerhouse. Based on her vita, letters of recommendation, and their own reading of her work, the committee rated X's research record as "extremely strong." X's recommenders all especially noted her high productivity, impressive analytical ability, independence, ambition, and competitive skills, with comments like "X produces high-quality research and always stands up under pressure, often working on multiple projects at a time." They described her tendency to "tirelessly and single-mindedly work long hours on research, as though she is on a mission to build an impressive portfolio of work." She also won a dissertation award in her final year of graduate school. X's faculty job talk/interview score was 9.5 out of 10. At dinner with the committee, she impressed everyone as being a confident and professional individual with a great deal to offer the department. During our private meeting, X was enthusiastic about our department, and there did not appear to be any obstacles if we decided to offer her the job. She said she is single with no partner/family issues. X said our department has all the resources needed for her research.*

As can be inferred from the above portion of text that was read by those faculty who were randomly assigned to this particular condition, the conditions systematically varied the lifestyle of the applicants. For example, some applicants, such as the one in the above example, were described as single without children, while others were described as married or divorced, with or without children, with or without a spouse who worked either at home or out of the home (to determine whether those making hiring decisions have implicit beliefs about the value of having "back-up" for childcare at home in the eventuality that school is closed or a child is too ill to attend). As noted, each version was paired with a mirror version in which the genders of Drs. X and Z were switched to compare identical candidates living under identical conditions, who differ only in their gender. This was done between-subjects to assess gender's influence, both as a main effect and as it interacted with lifestyle variables. As noted above, the faculty respondents were unaware of this, because each respondent received only one version and was not informed that a colleague elsewhere received the mirror version with Dr. X's and Dr. Z's genders reversed, but all other text remaining the same.

In sum, because the experimental design was "between-subjects", each faculty rater only received one set of three hypothetical applicants in one of the 20 experimental conditions to rank. Faculty respondents were unaware that other faculty were sent the identical three applicants they were sent with one critical change—the genders of the two strongest applicants, Drs. X and Z, were reversed; for instance, in the above example, other faculty received the identical description of Dr. X, but instead portrayed as a "he" rather than a "she". Thus, it was possible to compare the hirability rankings of two hypothetical applicants with identical records and identical 9.5 faculty ratings, who differ only in their gender (signaled solely by the presence or absence of the letter "s" in the pronoun (s/he). Across the 873 faculty respondents, half received Dr. X as a male and half as a female, and for every Dr. X depicted as a male gendered persona there was a counterpart depicted as a female persona. Within each of these groups some faculty received depictions of Dr. X as single, married, divorced, with or without children, etc.

Readers can consult the Resources section (http://www.ciws.cornell.edu) to see how family-related information was "smuggled" into the search chair's notes, with the explicit caveat that this information was spontaneously disclosed by the applicant, not solicited by the chair of the search—which would be strongly discouraged, of course. How widely known is family status in an era in which inquiring explicitly about marital status and children is discouraged? Our experience on search committees for tenure-track positions is that such information sometimes emerges during meetings and dinners with candidates. Many of our consultants confirmed that they, too, knew the marital status/children of some candidates in their departmental searches as

a result of the candidates asking about local schools, housing costs, and neighborhoods, work opportunities for partners, etc. Thus, while it is taboo to initiate discussions about such matters, sometimes candidates themselves divulge such information, opening the possibility to smuggle in family status this way in the search chair's notes. Was providing family information a problem for faculty respondents? Only 8 of the respondents raised this issue, informing us that although they appreciated that the search chair did not solicit marital/children information, at their institutions it is impermissible to directly request this information.

Experiments 1 and 2: Test of Effects of Lifestyle Difference—Comparing Candidates with Matching (Experiment 1) and Nonmatching (Experiment 2) Lifestyles. As mentioned, across the first three experiments there were ten lifestyles and two counterbalanced versions of each that were contrasted among otherwise identical applicants. This was done because an applicant's lifestyle is a potentially confounding variable in the assessment of hiring bias that has not been studied experimentally. Perhaps anti-female bias is more evident in response to certain lifestyles (e.g., a divorced mother with two preschool-aged children) than others (e.g., a single, childless woman or a woman who has a spouse who runs a home-based business who can help care for children in times of need). We included lifestyles suggested by past research as influencing women's relative likelihood of being hired, such as whether there is a marriage or child penalty (*5*). As was true of the test for gender, these contrasts of lifestyles were accomplished between-subjects so each faculty respondent viewed only one lifestyle, thus obscuring this contrast—except in conditions 7.1/7.2-8.1/8.2 which intentionally compared real-world lifestyle differences such as married father with stay-at-home spouse competing against a divorced mother. These cross-lifestyle comparisons were deliberately chosen to reflect potentially bias-inducing situations that may occur in the academy, given the demographics of the modern professoriate, in which more men are married with stay-at-home spouses than is true for women and in which more women are divorced mothers with custody of their children as compared to the numbers of divorced men with custody. Thus, 7.1/7.2-8.1/8.2 explored scenarios in which women might encounter bias.

Experiment 3: Test of Effects of Parental Leave. Parental leave is a topic of current national policy interest. Although a great deal has been written about the need for employment flexibility and work-life balance (6), there have been no experiments testing the effect of leave-taking on employment. Thus two of the lifestyle conditions examined the effect on applicants—fathers as well as mothers--of taking a one-year parental leave during graduate school. Does taking such a leave during graduate school affect identical male and female job candidates differently? If so, does taking a leave interact with the gender of the faculty member who is judging the leave-taker? In order to avoid confounding gender and parental leave, these last two conditions always pitted a male Dr. X against a male Dr. Z (with Dr. Y being a female foil) or a female Dr. X against a female Dr. Z (with Dr. Y being a male foil). Thus, unlike in the other conditions in the first three experiments, these two conditions were same-sex contrasts, not cross-sex contrasts. The sole difference between these applicants was whether they took a one-year parental family leave while working toward their Ph.D. Some faculty respondents received Dr. X as the leave-taker and others received the same-sex Dr. Z as the leave-taker, thus permitting a comparison of otherwise identical applicants who differed only on the leave-taking variable. These contrasts also provided a test of whether certain gendered personae are favored in hiring decisions, as we discuss next.

II. Validity and Generalizability Issues

a. Validity Checks on Sample Response Bias. Email experiments usually report response rates in the 25%-35% range (e.g., 7, 8), thus introducing the risk of biased samples. A means of validating the assertion that the respondents represent an unbiased estimate of the underlying population from which they were drawn is to examine whether demographic variables in the sample of respondents (e.g., sex, rank, discipline) approximate their counterpart in the overall population that was sampled. When demographic variables in the sample approximate the overall population, survey research has suggested that respondents tend to be similar to non-respondents in their responses to focal variables (e.g., 8, citing chapter by 9). Examination of the sample characteristics in this study indicated that the 34% of respondents reflect those of the population from which they were drawn in terms of their gender, rank, and discipline. In the past, this has often been the only validation check researchers have utilized in experimental email surveys. However, in the present research we endeavored to go beyond this form of validation.

Notwithstanding the above form of validation, it is still possible that the 34% of respondents differed from the 66% of non-respondents in unmeasured attitudes regarding gender equality and that these attitudes are nonorthogonal to one or more of the measured variables. Therefore, we sought to provide a stronger form of sample validation. To accomplish this, we undertook two additional checks. In the first of these, we offered a $25 gift card to 90 faculty members drawn from the same pool of institutions as the overall sample from one of the four fields—psychology; 91.1% of these psychologists (82 out of 90) provided full response data. Without exception, every analysis of the distribution of responses from this 91.1% paid sample did not differ from the distribution of psychologists' responses in the 34% sample in the main experiment. Taken together, the results of these two validity checks suggest that the 34% sample is an unbiased estimate of the underlying population.

Sample Weights. However, we also undertook another validity check that is even stronger: In addition to analyzing the respondents' data using a traditional statistical approach, we also analyzed it using sample weights to adjust for any form of nonrandomness that could limit generalization. Below we describe the rationale for using weights and the calculated weights used in this analysis. As will be seen, these weighted analyses resulted in very similar results to the unweighted ones reported in the main article. Every finding reported in the unweighted analysis in the main text was confirmed in the weighted models. (Interested readers should consult the Resources section of this Supplement for a detailed statistical report and for an independent statistician's replication of the findings from the first experiment, at: www.ciws.cornell.edu.) Taken together, these three efforts at validity-checking gave us confidence that the findings generalize beyond the specific sample to the population of Carnegie institutions in the sampling frame. This covers all institutions of higher education, save special focus ones (e.g., seminaries, culinary institutes, stand-alone business schools that are not part of larger institutions, tech training colleges, etc.) as long as they have three or more of the four fields of interest.

Weighting respondents according to their probability of being selected for inclusion in a sample is a germane consideration when the probability of randomly sampling a given woman in a male-dominated field is far higher than the probability of sampling one of her male colleagues because there are fewer women to choose among. When there are fewer women in a given department (as is the case in economics and engineering), then randomly selecting one male and one female from these departments will mean that the odds of selecting any particular woman is higher than the odds of selecting any one of her male colleagues. Because of this, any given woman's data "stands in" for fewer unsampled female colleagues (but a higher proportion of them) than does any given man's data. Hence the woman's data can be weighted to reflect this fact.

One way to deal with this is to construct a two-stage statistical model, weighting both the probabilities of randomly selecting a particular institution as well as the probability of randomly sampling a particular man and woman in that institution. This is because certain types of institutions, such as doctoral-granting ones, may have different representations of women than, say, small colleges, and their faculty may also have different attitudes about the value of hiring women. In the first stage of a two-stage model, weights are calculated for the likelihood of sampling each type of institution (Doctoral Intensive, Combined Master's/Baccalaureate institutions, and Baccalaureate institutions) from the universe of potential institutions of this sort in the country, and these weights are then multiplied by the odds (or inverse of odds) of randomly sampling a woman and man within this institution from their home department within each type of institution. To accomplish this department-based weighting, one needs to know how many men and women are in any given respondent's or nonrespondent's home department. In the second stage of the model, post-stratification weights can be added to adjust for differences in strata between respondents and nonrespondents, e.g., male economists at mid-sized institutions who responded vs. did not respond.

There are pro and con arguments regarding the usefulness of weighting at neither, one or both stages. The topic has generated substantial discussion, with some favoring one approach to creating weights and others favoring another approach. Moreover, some readers (from disciplines in which data tend not to be manipulated or transformed) may react skeptically to such transformations (e.g., weighting men's data more heavily than women's in the analyses may strike some as moving farther away from the raw data, thus engendering skepticism). For example, while weighting acknowledges the differences between sampling a particular woman versus a particular man in an engineering program that has, say, 80% male faculty—and then adjusts the collected data to reflect these underlying sample characteristics—weighting can also skew the results in

terms of the likelihood that a given applicant will be hired, because each faculty member, regardless of gender, has only a single vote, which weighting can obscure. Consequently, we undertook complete separate analyses of our data both with and without weighting. The results reported in the main text are derived from non-adjusted (unweighted) logistic models and those reported in this supplement are weighted at stage 1 but not at stage 2 (more on this below). As will be seen, the results were essentially the same in the weighted and un-weighted analyses: none of the findings reported in the main text changed when the data were weighted.

As noted, we oversampled doctoral institutions 2:1 compared to master's and baccalaureate institutions. This was done for various reasons: First, doctoral-granting universities are the premier institutions, with the most stable and well-paying faculty posts (http://www.aaup.org/sites/default/files/files/2014%20salary%20report/Table4.pdf); we wanted enough statistical power to do stand-alone analyses of these institutions because we suspected that however the findings came out many would be most interested in whether they were limited only to teaching-intensive jobs that are not as highly remunerated, or whether they applied to R1 jobs as well. Second, doctoral-granting universities are also the institutions in which women faculty have historically been highly underrepresented in engineering and economics (fewer than 15% of all faculty). Third, doctoral institutions are among the largest, accounting for a disproportionately large portion of both students and tenure-track faculty. Finally, whereas most of the master's and baccalaureate institutions do not have one or both of these math-intensive fields, all doctoral institutions contain all four fields. Thus, we aimed from the beginning of the study to sample twice as many of the nation's doctoral programs as its master's and baccalaureate ones, a goal that was achieved. Among all respondents who returned complete data, 57% were employed at doctoral institutions.

Weighting was carried out in two steps. The three levels of the *Carnegie Basic Classification* scheme is based on quantitative analysis of such variables as the selectivity of the student body, enrollment and graduation rates, and institutional R & D expenditures. The three levels of the *Carnegie Basic Classification* (Doctoral Intensive, Combined Master's/Baccalaureate institutions, and Baccalaureate institutions) are further subdivided according to their size and research-intensity. For example, there are small (S), medium (M) and large (L) Combined Master's/Baccalaureate institutions; among Doctoral institutions there are those with very high research activity (RU/VH), high research activity (RU/H), and others (DRU); Baccalaureate colleges have three levels that are based on their focus and degree to which they are involved in associate (2-year) degree training. Table S.2 shows the breakdown of these institutions with their national numbers, the fraction they comprise of all institutions, and their mean student body sizes.

Table S2. Absolute number and percentage of institutions and student enrollment for doctoral, master's and baccalaureate institutions.

	Total # Institutions	%	Total Enrollment	%	Average Enrollment
RU/VH: Research Universities (very high research activity)	108	2.3 %	2,809,581	13.6%	26,015
RU/H: Research Universities (high research activity)	99	2.1 %	1,746,651	8.4 %	17,643
DRU: Doctoral/Research Universities	90	1.9 %	1,228,846	5.9 %	13,654
Master's L: Master's Colleges and Universities (larger programs)	413	8.9 %	3,503,396	16.9%	8,483
Master's M: Master's Colleges and Universities (medium programs)	185	4.0 %	785,985	3.8 %	4,249
Master's S: Master's Colleges and Universities (smaller programs)	126	2.7 %	367,219	1.8 %	2,914
Bac/A&S: Baccalaureate Colleges--Arts & Sciences	271	5.8 %	460,036	2.2 %	1,698
Bac/Diverse: Baccalaureate Colleges--Diverse Fields	392	8.5 %	664,939	3.2 %	1,696
Bac/Assoc: Baccalaureate/Associate's Colleges	147	3.2 %	298,300	1.4 %	2,029

Source: Carnegie Foundation for the Advancement of Teaching, 2011.
http://classifications.carnegiefoundation.org/summary/basic.php

Pre- vs. Pre- and Post- Weighting. First, we calculated the probability that a given institution was randomly selected for inclusion in the study. This calculation was straightforward for Doctoral institutions but not for Combined Master's/Baccalaureate institutions or for Baccalaureate-only ones. There are 297 doctoral institutions (108, 99, & 90, for RU/VH, RU/H, and DRU, respectively), and the probability of any one of them being sampled was 60.1% (181/297).

The probability of any one of the 724 Combined Master's/Baccalaureate institutions being sampled had to be adjusted because: a) over half of them either do not contain at least three of the four fields (e.g., University of Colorado at Pueblo does not have programs in psychology and economics), b) they were special-focus institutions that do not have tenure-track faculty in the fields in question (e.g., seminaries, schools of professional education, business administration), or c) they were vocationally-oriented, awarding professional degrees in fields such as criminal justice, medical technology, human services, dental hygienics, veterinary technical services, bible studies, etc., but they awarded no degrees in the four fields of this study (e.g., Argosy University, Columbia International University). As an illustration of the need to exclude the majority of Combined Master's/Baccalaureate institutions, Table S3 in the Resources section (http://www.ciws.cornell.edu) shows the first 50 institutions from the category of large Combined Master's/Baccalaureate programs. As can be seen, 32 of these 50 were ineligible for inclusion for one or more of the reasons above (asterisked). After deleting 416 of such institutions, the probability of sampling one of the remaining 308 was 27.3% (i.e., 84 out of 308). This was within our target of sampling 50% of doctoral institutions and 25% of the other two types.

Table S3. Carnegie Foundation's first 50 large master's program, of which 32 do not have programs in at least 3 of the fields. Asterisked (*). See SI Resources at: http://www.ciws.cornell.edu

As an example, consider Baccalaureate colleges, which number 807. However, most of these were also ineligible for inclusion in the sample either because they are special emphasis institutions (e.g., seminaries, culinary institutes, Vo-Tech campuses), or because they are primarily institutions that confer 2-year associate degrees, or do not have tenure-track faculties in at least three of the four relevant disciplines. Only 170 of the 807 Baccalaureate colleges contain at least three of these four fields. The probability of one of these institutions being sampled was 34.1% (58/170). Overall, there was a total possible pool of 775 institutions that met the inclusion requirements, 452 of which we sampled, and 347 of which contributed data to Experiments 1-3. Within each of these 452 institutions, anywhere from 6 to 20 male and female faculty members were randomly sampled from 3 to 4 departments, starting with one male and one female randomly selected in each of at least three disciplines, with replacements randomly chosen for use in the event of nonresponse, and additional faculty selected for various validation studies. Of the 2,090 faculty randomly sampled from the pool of 452 eligible institutions across all three Carnegie categories, 711 faculty responded with full data. As seen below, the response rates were fairly similar across the three categories of Carnegie institutions, although there was a range across the four disciplines. (The paid validity check was conducted only on psychology faculty, hence the response rate for that discipline was the highest; however, aside from this paid validation sample, the response rate among psychologists was similar to the response rates for the other three fields.) The probability of a given institution being sampled was calculated: this is the fraction of the number of unique institutions from which respondents came divided by the total number of eligible institutions in that Carnegie classification category.

For each of the 711 faculty who returned data, the probability that a given male or female respondent in their institution was sampled in their department was calculated from the numbers of female/male tenure-track faculty in it; this was done for every respondent in each of the four fields at each of the 347 institutions. Thus, the inverse of these two values—the probability of a given institution being sampled and the probability of the individual being sampled within it—were multiplied to calculate weights for each of the respondents. As seen in the table below, because of their relatively smaller fraction of the respondent pool for the six conditions of the main experiment (Total N=363; analyzed N=339) female faculty "stood for" 2.0 to 2.5 times more same-sexed peers than did their male counterparts, and therefore their weights were reduced accordingly in a second set of logistic regression analyses (see Table S4).

Carnegie Category	Field	Faculty Gender	Total #	Total # Inst. in Sample	Ind. Weight	# Inst.	# Inst. Respond	Inst. Weight	Ind. X Inst. Weight	Total of Weights	Normalized Weights
1	Bio	Female	471	51	9.2353	297	194	1.5309	14.139	721.07	0.6256
1	Bio	Male	1585	73	21.7123	297	194	1.5309	33.24	2426.52	1.471
1	Econ	Female	257	37	6.946	297	194	1.5309	10.634	393.449	0.471
1	Econ	Male	973	44	22.114	297	194	1.5309	33.854	1489.59	1.498
1	Eng	Female	252	60	4.2	297	194	1.5309	6.4299	385.794	0.285
1	Eng	Male	1133	55	20.6	297	194	1.5309	31.537	1734.54	1.395
1	Psy	Female	1072	86	12.465	297	194	1.5309	19.083	1641.16	0.8444
1	Psy	Male	1413	76	18.592	297	194	1.5309	28.463	2163.20	1.259
2	Bio	Female	117	19	6.1579	308	95	3.2421	19.965	379.326	0.8834
2	Bio	Male	179	14	12.786	308	95	3.2421	41.453	580.337	1.834
2	Econ	Female	62	13	4.77	308	95	3.2421	15.462	201.011	0.6842
2	Econ	Male	82	9	9.1111	308	95	3.2421	29.539	265.853	1.307
2	Eng	Female	49	14	3.5	308	95	3.2421	11.347	158.863	0.5021
2	Eng	Male	161	12	13.4177	308	95	3.2421	43.498	521.989	1.9247
2	Psy	Female	205	27	7.593	308	95	3.2421	24.616	664.632	1.0892
2	Psy	Male	253	28	9.03571	308	95	3.2421	29.295	820.253	1.2962
3	Bio	Female	43	8	5.375	170	58	2.9310	15.754	126.035	0.6971
3	Bio	Male	38	8	4.75	170	58	2.9310	13.922	111.379	0.6160
3	Econ	Female	62	16	3.875	170	58	2.9310	11.358	181.724	0.5026
3	Econ	Male	116	14	8.2857	170	58	2.9310	24.286	340.0	1.0746
3	Eng	Female	12	3	4.0	170	58	2.9310	11.724	35.172	0.519
3	Eng	Male	54	6	9.0	170	58	2.9310	26.379	158.276	1.167
3	Psy	Female	21	98	4.6667	170	58	2.9310	13.678	287.241	0.6052
3	Psy	Male	96	17	5.6471	170	58	2.9310	16.552	281.379	0.7324

Table S4. Sample weights (Carnegie Category: 1=doctoral institutions; 2=master's/baccalaureate institutions; 3=baccalaureate colleges).

In addition to this "front-end" weighting for each respondent, we considered incorporating a post-stratification weighting for each of the 1,379 nonrespondents, similarly calculated. Such weights are the inverse of the selection probability, with adjustments for nonrespondents in each strata (e.g., male engineers at doctoral institutions). There are advantages and disadvantages to using post-stratification weights (see 10 for discussion, p. 364). After consultation with five statisticians, including Martin T. Wells, coauthor of the article cited above, we decided against adding this second set of weights. Instead, as recommended by Wells, we conducted a logistic regression analysis of the difference between two 24-cell data sets of all combinations of strata (2 gender x 3 Carnegie types x 4 fields), comparing respondents and non-respondents. (Note that we first omitted the 82 paid psychology respondents since their response rate of 91.1% was inflated due to the compensation offered.) Any non-randomness costs should be revealed in differences across strata. The comparisons of the strata across these two groups were similar in every way. For example, the ratio of respondents to nonrespondents was statistically similar for men and women, and this remained true in each of the four disciplines and at each of the three Carnegie institution categories. In sum, coupled with the random sample of 82 tenure-track psychologists who were remunerated to participate, resulting in a 91% response rate (82 out of 90), the weighted analyses suggest the data in the 34.02% sample are not different from what would be expected as one approaches full population buy-in by respondents from the pool of 775 eligible institutions.

Study-wide Power. Study-wide power concerns maximizing total statistical power across multiple experiments (11). Total power is most informative for studies that involve multiple experiments and multiple

hypothesis tests. Because power decreases as the number of experiments/tests increases, significance testing across experiments can be calculated for the focal hypothesis of whether a sex preference is exhibited by men and women faculty, both overall and for the two math-intensive fields when combined. Experiments 1, 2 and 4 all involve the focal hypothesis of 2 (gender of faculty) x 2 (gender of applicant) tests of interactions, collapsing across the four fields as well as combining the two math-intensive ones vs. the two non-math-intensive ones, which enlarges samples and confidence that the headline finding is valid—women are preferred roughly 2:1 over identically-qualified men, both overall and when the combined two math-intensive fields are pitted against the combined two non-math-intensive fields. For 80% power across these three between-subjects experiments (alpha=.05, two-tailed), a sample of 228 is needed to detect large effects (11). For Experiments 1, 2, and 4 (which all contain the focal gender x gender interaction, collapsed across fields), our sample was 542, easily capable of detecting large effects, which others have reported with regard to a similar question (8). (For additional information on the focal hypothesis testing, see *Resources* at: www.ciws.cornell.edu.)

b. Data about Academic Disciplines. As already noted, 711 respondents in the first three experiments returned data, with the percentages of the response pool being similar in 3 of the 4 disciplines: psychologists (35.9%), economists (18.7%), engineers (21.1%), and biologists (24.3%). Again, the reason for the higher proportion of psychologists (35.9%) is because of the paid validation test described above in which respondents were offered a $25 gift card for participating, which was conducted only in the field of psychology and which resulted in a 91% response rate for the paid subsample, thus increasing psychologists' numbers and elevating their overall representation in the sample. Excluding this subset of paid psychologists, the representation rates across the four fields were roughly comparable: biology (27.5%), economics (21.1%), engineering (23.8%), and psychology (27.5%).

c. Validation of Narrative Research Summaries vs. CVs (Experiment 4). Initial materials and the full experimental design were reviewed and revised based on feedback from national consultants to this project--senior faculty from these academic fields. Their feedback guided the development of the descriptions of the three hypothetical applicants and the cover materials sent to respondents. These materials, once revised, were deemed feasible and veridical for each field by the relevant experts in it. Initially, identical CVs were prepared for the candidates in each discipline. However, pilot testing led to abandoning their use (as well as replacing male and female names with letters X, Y, and Z thereafter described with male/female pronouns, to avoid any gendered connotations that a given name might hold). The reason for the abandonment of CVs follows.

Instead of using identical CVs for the top applicants in each discipline, narrative summaries of their research achievements/interview ratings were substituted, with accompanying numerical scores on a 10-point scale, in which 10=truly extraordinary/exceptional and 1=cannot support. This decision to abandon using identical CVs for the top male and female applicants was made because pilot testing revealed that institutions differed substantially in what they considered to be an extraordinary/exceptional applicant in terms of the quantities and types of publications needed to be short-listed for an entry-level tenure-track position. In fact, sometimes even subfields within a field differ, as for example, in the value placed on published proceedings in electrical engineering versus mechanical engineering, according to two consultants on this project. Within our own field of psychology, subfields differ vastly in terms of typical publication counts for successful applicants for top-tier jobs, with applicants in social psychology often possessing many more publications than those in cognitive psychology and three times as many as successful applicants in developmental psychology (12).

Thus, the primary reason for abandoning identical CVs is that universities differ greatly in what is regarded as an impressive or unacceptable candidate, with highly selective PhD-granting institutions expecting more publications, eminent advisors, awards earned in graduate school, etc., than do many small teaching colleges. No single CV was deemed by our national consultants as a realistic portrayal of an impressive applicant across all types of institutions, spanning doctoral-intensive universities to 4-year teaching-intensive colleges. Attempts to individualize CVs for these fields and institutions introduce non-comparability problems for analyses that rely on collapsing data across fields and institutions: a short-listed applicant who had 6 peer-reviewed journal publications at one institution was not necessarily comparable to a short-listed applicant in another field or institution who may have fewer publications, or in lieu of any peer-reviewed articles a book based on the dissertation (more common in economics than in engineering, for example). Or, to use another example of potential non-comparability, an applicant to an elite institution may be expected to have an advisor who is eminent—mentioned by two consultants as an important consideration in hiring in their own depart-

ments—whereas this is not expected of an applicant from another field or to most institutions). For these reasons CVs were abandoned and narrative statements were substituted that allowed each faculty at each institution to assess applicants in the context of their own needs and expectations. We left it to them to contextualize what is meant by a 9.5 (excellent) rating in their own department searches in terms of the number and type of publications they regarded as excellent rather than itemizing publications.

In sum, the use of narrative research summaries rather than itemized publications acknowledges the relative nature of institutional searches. What is considered an impressive candidate at most highly select research-intensive doctoral degree-granting institutions is different from the judgment made at many master's-degree institutions and small teaching-intensive colleges. Thus, to avoid conceptual and statistical problems inherent in sending out a CV that had an impressive number of publications for a short-listed candidate at a research-intensive institution but perhaps an unrealistically high number of publications for some teaching-intensive institutions, we instead sent summaries that used phrases to signal research activity (e.g., "Everyone agreed she is building an impressive research record") rather than absolute numbers of publications, so that faculty from each institution could place these phrases in their own context.

Our national consultants helped us design a common 10-point research rating scale that would be understandable and feasible for faculty in their fields. As can be seen in the Resources section (online), our summaries contained ratings by the hypothetical departmental faculty for the top three applicants, Drs. X, Y and Z, with X and Z counterbalanced for gender, and Y being a slightly-lower-rated male foil (except in Condition 10 which had a female foil because Drs. X and Z were both males), chosen to divert attention from a male-female match-up between X and Z. On this 10-point scale, Dr. X's and Dr. Z's research quality and productivity were both rated as 9.5 (which corresponded to a rating of "impressive" on the scale faculty respondents were given), without mentioning specific numbers or types of publications that went into the decision to rate the applicant 9.5. (As noted above, Y was rated slightly below them, at 9.3, and was included to obscure the head-to-head competition between a male and a female.) Thus, the cover instructions sent to each faculty respondent stated that their departmental colleagues had already reviewed the actual CVs and job talks and interview ratings and gave these applicants an overall rating of 9.5 (or 9.3 in Y's case), using common descriptors such as "The search committee rated Z's research record as 'extremely strong'", and allowing each respondent to interpret what this means in the context of their own department. (See examples in Resources section.)

Are these narrative summaries satisfactory substitutes for CVs for the purposes of ranking of applicants? As a check on their validity, we asked a group of 35 professors of mechanical engineering at Carnegie-classification Ph.D.-granting institutions--drawn from the same sampling frame used in the main study--to perform the same ranking task. But instead of using narrative research summaries, we gave them three full CVs, with the CVs for X and Z containing one more publication each and one more presentation at a conference each than was true for the CV for Y, to approximate the difference between 9.5 and 9.3 ratings (i.e., the difference between the two outstanding candidates and strong foil). (All three of these CVs were based on those of actual new Ph.D.s in mechanical engineering from Ph.D.-granting Carnegie-classification institutions that we downloaded from the web and doctored to fit the needs of this validation study while obscuring the identities of those on whom they were based.) One of our national consultants in engineering, an endowed-chair holder and the co-PI on a 100-million-dollar NSF grant, informed us of the different expectations in the engineering subfields in terms of whether conference proceedings were highly regarded, and urged us to stick within a single subfield of engineering to avoid non-comparability issues when using CVs. Everything else about this task was identical to the task in our overall experiment, and these engineering professors were asked to make the same rankings undertaken by their colleagues who were given narrative summaries.

As expected, the distribution of the rankings of this group of engineering faculty, 19 men and 16 women, was not significantly different from that of the engineering professors in the main sample who based their rankings on the narrative research summaries: every statistical comparison supported this interpretation. In fact, the raw tallies when actual CVs were used revealed that the woman candidate was chosen over the identical man candidate by an even larger margin than in the main study (75.8% of engineers chose the woman in the full CV experiment vs. 66.7% of engineers chose the woman in the main study), although this difference was not significant. This confirms that research summaries were suitable proxies for CVs and showed comparable female preference, while, unlike CVs, having the advantage of comparability across institutions, fields, and subfields.

d. Individual vs. Group Decision-Making. In virtually all experimental studies of sex differences in both academic and nonacademic hiring, researchers have used variants of the design we used here (see meta-analysis *19*). Past researchers gave raters the same applications/CVs but varied the gender of the applicant's name on them. Despite this being the method used in most past research, however, academic hiring decisions are usually made by committees or by the entire departmental faculty, rather than by tallying individual votes made in private and without public discussion. Therefore, one might wonder if the group versus individual deliberation process is responsible for the strong female preference found here. Perhaps this female preference would be mitigated if faculty were instructed to engage in public discussion before making public rankings. Would public group decision-making have resulted in a less strong pro-female preference, perhaps even a pro-male preference? Before addressing this question it is worth asking why a private individual ranking would favor females over males while a public discussion would not. And independent of the group-vs.-individual distinction, how does one explain the 2-to-1 preference for females in the present individual condition when the opposite is usually found in studies that used similar individual condition?

Research suggests that, if anything, a public discussion and decision-making can be expected to result in an even stronger pro-female preference than what would be found in private decision-making. The social psychological evidence reveals that asking groups of individuals to collaboratively decide on whether to hire a female (or minority group member) will actually lead to higher rates of female or minority preference, not lower rates. It might even be the case that group decisions would exceed the 2-to-1 ratio found for private individual ratings. At the very least there is no compelling evidence to suggest that group processes would have resulted in an anti-female bias, and there is anecdotal as well as scientific evidence that group discussion would, if anything, reduce bias against women and minorities. Anecdotally, researchers noted that in their department's effort to increase female and minority hiring: "*Simply discussing unconscious bias heightened our sensitivity*" (13, p. 612). As already noted, past experiments that alleged anti-female bias in hiring used the same individual private ratings as employed here, suggesting this feature is capable of revealing bias if and when it exists. Below we elaborate on the scientific evidence for the claim that, if anything, public decision-making would not have led to a male preference and possibly even have led to a greater female preference.

Research shows that people are aware of the social norms against expressing prejudice; in public settings they suppress the expression of bias, or express it in more subtle ways (14-16). We know of no evidence that fewer females (or minorities) would be preferred in a public setting in today's academic environment than in a private setting in which individuals cast confidential votes, as they did in the present experiment. Moreover, research suggests that private behaviors, such as the preference ratings used here, are often quite predictive of future behavior (15, 16). This classic work shows that behavior is a function of attitudes, intentions, and social norms, such as trying to enhance diversity. This may explain why experiments in which individuals consciously or subconsciously reveal their true feelings and make judgments based upon them are more successful in revealing bias than are analyses of actual hiring. Actual hiring may involve lengthier study of application materials and discussion, as well as actual face-to-face interaction that challenges stereotypes and keeps biased raters "on guard" and "on their best behavior." In sum, everything points in the direction of even stronger female preference if the decision-making is public. However, independent of this issue we are still faced with the need to explain why there was a 2-to-1 preference for women when the rankings are done privately, given the frequent contrary claims based on the same methodology. We return to this question below but first describe several validation studies.

e. Additional Empirical Validity Check Regarding Gendered Personae (Adjective-sex). The adjectives chosen to describe the candidates were derived from studies evaluating the prototypicality of various adjectives on the male-to-female dimension, such as "ambitious, analytic, powerhouse" on the male end vs. "kind, creative, socially-skilled" on the female end (*3, 4*). Such prototypicality judgments are usually made by students and laypersons, so it is of interest to document that faculty view them similarly. We validated these adjectives by asking 50 faculty (chosen from the same sampling plan that generated the 363 faculty whose data formed the basis of the first experiment) to rate the stimuli used in the experiment as descriptors for two different personae—female and male. We asked them to rate these descriptors on a scale from male to female, using intermediate numbers if desired: 1=Prototypical Male Persona, 3=More Male than Female Persona, 5=Neutral Gender Persona, 7=More Female than Male Persona, 9=Prototypical Female Persona. Of the 50 faculty asked to do this, 33 provided data (66%). As expected, the concordance was high: 31 of the 33 faculty

rated the adjective toward the pole of the scale that was expected ($p<.0001$ by one-tailed sign test). Looking at the individual scales, the mean faculty rating of the male gendered persona was 2.58 and the mean faculty rating of the female gendered persona was 6.70, $t(31)=11.76$, $p<.0001$, SE=.34. Thus, the adjectives that have previously been shown to be associated with gendered personae in the general population were confirmed among a sample of faculty.

III. Procedures

a. Sample Demographics. We began by compiling the complete list of colleges and universities from the Carnegie Basic Classification list of all Code 3 (various forms of Baccalaureate colleges), Code 2 (various forms of Combined B.A./Master's institutions), and Code 1 institutions (research-intensive universities that grant a certain number of PhDs in a threshold number of fields). These three codes do not include stand-alone institutions such as seminaries and tech centers that would not be relevant to our hypothesis; however, they also exclude stand-alone medical schools that are relevant, so we included the latter in our full sample of institutions. From the full list of Carnegie 1, 2, and 3 institutions and medical schools, we oversampled Carnegie code 1 institutions (research-intensive universities) because these were of special interest to us; they contain the premier academic positions in terms of salary and training, they cover a disproportionately large portion of students and faculty, and they disproportionately constitute the top 100 or so universities in the *U.S. News & World* rankings. These institutions often have pronounced underrepresentations of women in math-intensive fields. Thus, our goal from the outset was that approximately half of all institutions would be from these research-intensive universities and the other half divided between Code 2 and Code 3 institutions.

Within the three types of institutions, we sampled four departments (engineering, psychology, economics, and biology), randomly choosing one male and one female faculty member from each of these four departments and replacement faculty members in the event the original choices failed to respond, plus additional faculty for the validation studies. After deleting inaccurate/bounceback emails and people who informed us they had retired, were not tenure-track, etc., and replacing them with randomly-selected substitutes from the same departments, we were left with 2,090 potential respondents, of whom 711 contributed data to the first three experiments (response rate=34.02%). These 711 respondents reflected our sampling plan, thus coming from Carnegie 1 research-intensive institutions just over half of the time (56%), with the balance of respondents divided between Carnegie 2 and 3 institutions. As noted earlier, we augmented this sample with 35 additional faculty to validate the use of narrative summaries versus CVs, bringing the total sample to 746, and we further augmented it with 127 faculty to examine the rating of each sex when faculty were asked to evaluate only a single applicant, either a male or a female, bringing the total to 873. Finally, 33 faculty completed the gendered-adjective rating task above.

We coded and entered the Carnegie codes for each of the 711 respondents' colleges/universities into an Excel file. (Note, this excluded the 35 engineers from Experiment 4, the 127 respondents in the Experiment 5 modified task, and the 33 faculty who did the gendered-adjective task.) We also created a separate Excel file containing the Carnegie code data on each of the 1,379 nonrespondents. Next we constructed a 24-cell table: 2 genders x 4 departments x 3 Carnegie codes. For example, there is a cell for female engineers from Carnegie Code 3 universities, and there is a cell for male biologists from Carnegie Code 1 colleges/universities, etc. All 24 cells have the frequencies and ratios of respondents-to-nonrespondents. This permits us to determine whether, say, female engineers from Code 1 colleges/universities were more likely to respond than were males or females in any of the four departments in any of the Carnegie code categories. As an example, suppose that we emailed 60 female engineers from Code 3 institutions and only 20 responded with data. In that cell (female, engineer, Code 3) would show 0.333 (20/60). Of the 711 respondents in the first three experiments who returned full data, these response rates were fairly similar across fields. And, as noted in the main text, men and women were nearly equally represented among respondents.

b. Additional Information on Three-Part Research Design. In experiments it is common to systematically change the levels of one variable at a time. This set of experiments varied one thing at a time in a between-subjects' design, which was the gender of pronoun used to describe otherwise-identical applicants. If the only thing that was varied within each condition was the gender of the applicant for a given faculty rater, faculty raters might be able to detect the underlying purpose of the experiment and behave artificially. That is, pitting a man against a woman who are in all other respects identical makes the gender contrast salient and

might influence raters' behavior. In particular, it might lead to politically-correct responding in which faculty who might otherwise exhibit bias against females rate them higher because they realize their choice is being scrutinized. Therefore, the central hypotheses were tested using a three-part between-subjects design so that the gendered nature of the hypothesis was obscured. Each part of this design targeted a different hypothesis, as explained below. And, as already noted, two ploys were introduced to mask the gender contrast: first, the use of a slightly weaker male foil, and second, the use of different gendered personae (achieved by varying stereotypical sex of adjectives used to describe women and men with identical scholarly ratings) that were counterbalanced across gender of applicants, leading a sample of respondents to opine that the purpose of the experiment was to assess their relative preference for kind, creative, socially-skilled applicants versus analytical, ambitious powerhouses. In fact, none of the 30 faculty surveyed guessed that the real purpose of the study was assessing their gender preference. (Incidentally, had faculty realized the true nature of the study, the obvious option would have been to rank the two highest-rated applicants with a tied vote rather than rating the female higher, but this was done by only 3.5% of respondents.)

Experiment 1 was the core of the experiment. It was a fully counterbalanced contest between identical male and female applicants across various lifestyles that all applicants shared (e.g., single without children, married with children, divorced with children). This experiment crossed the applicants' gender with male and female gendered personae to obscure the gendered nature of the hypothesis; male versus female applicants were depicted with male versus female adjectives equally often, between subjects. This first experiment is the core of the project inasmuch as it provided comparisons of identical candidates with identical lifestyles differing solely in gender pronoun used to refer to them, rated by different faculty. Thus, both of these candidates who were rated by a given faculty member shared their lifestyle status and their accomplishments. Below are the six conditions in Experiment 1, each counterbalanced (thus 12 total versions), which permitted a direct statistical test of whether identical male and female applicants fared equivalently:

Experiment 1: First Six Crossed Conditions:

Total N for these first six conditions = 363 faculty, 182 female respondents and 181 male respondents.

1.1 Dr. X=woman described in female adjective condition, single no kids; Dr. Y=male foil; Dr. Z=man described in male adjective condition, single no kids.

1.2 Dr. X=man described in female adjective condition, single no kids; Dr. Y=male foil; Dr. Z=woman described in male adjective condition, single no kids.

2.1 Dr. X=woman described in female adjective condition, married no kids; Dr. Y=male foil; Dr. Z=man described in male adjective condition, married no kids.

2.2 Dr. X=man described in female adjective condition; married no kids; Dr. Y=male foil; Dr. Z=woman described in male adjective condition; married no kids.

3.1 Dr. X=woman described in female adjective condition; stay-at-home husband with kids; Dr. Y=male foil; Dr. Z=man described in male adjective condition; stay-at-home wife with kids.

3.2 Dr. X=man described in female adjective condition; stay-at-home wife with kids; Dr. Y=male foil; Dr. Z=woman described in male adjective condition; stay-at-home husband with kids.

4.1 Dr. X=woman described in female adjective condition; home-business husband with 2 kids in daycare; Dr. Y=male foil; Dr. Z=man described in male adjective condition; home-business wife with 2 kids in daycare.

4.2 Dr. X=man described in female adjective condition; home-business wife with 2 kids in daycare; Dr. Y=male foil; Dr. Z=woman described in male adjective condition; home-business husband with 2 kids in daycare.

5.1 Dr. X=woman described in female adjective condition; husband attorney with 2 kids in daycare; Dr. Y=male foil; Dr. Z=man described in male adjective condition; wife doctor with 2 kids in daycare.

5.2 Dr. X=man described in female adjective condition; wife attorney with 2 kids in daycare; Dr. Y=male foil; Dr. Z=woman described in male adjective condition; husband doctor with 2 kids in daycare.

6.1 Dr. X=woman described in female adjective condition; single mother (no ex-spouse in town) with 2 kids in daycare; Dr. Y=male foil; Dr. Z=man described in male adjective condition; single father (no ex-spouse in town) with 2 kids in daycare.

6.2 Dr. X= man described in female adjective condition; single father (no ex-spouse in town) with 2 kids in daycare; Dr. Y=male foil; Dr. Z=woman described in male adjective condition; single mother (no ex-spouse in town) with 2 kids in daycare.

Experiment 2 was an extension of the first experiment that pitted two of the lifestyles against each other in a direct contest. Thus, conditions 7.1/7.2 and 8.1/8.2 were designed to examine two targeted hypotheses: 1) whether a male applicant with a stay-at-home wife and two children is favored over an identically-qualified divorced mother with two children in daycare (Conditions 7.1/7.2), or 2) whether a male applicant with a stay-at-home wife and two children is favored over an identically-qualified single woman without children (Condition 8.1/8.2).

Total N for Conditions 7 and 8 = 144 faculty respondents, 80 males and 64 females.

Total N for Condition 7 = 69 faculty respondents, 39 males and 30 females.

Total N for Condition 8 = 75 faculty respondents, 41 males and 34 females.

7.1 Dr.X=man described in female adjective condition, 2 children and stay-at-home wife; Dr.Y=male foil; Dr.Z=woman described in male adjective condition, divorced with 2 children and absent ex-spouse.

7.2 Dr.X=man described in male adjective condition, 2 children and stay-at-home wife; Dr.Y= male foil; Dr.Z=woman described in female adjective condition, divorced with 2 children and absent ex-spouse.

8.1 Dr.X=man described in female adjective condition, 2 children and stay-at-home wife; Dr.Y=male foil; Dr.Z=woman described in male adjective condition, single with no children.

8.2 Dr.X=man described in male adjective condition, 2 children and stay-at-home wife; Dr.Y=male foil; Dr.Z=woman described in female adjective condition, single with no children.

Experiment 3 was a direct test of the impact of taking a 1-year parental leave during the final year of graduate school. Unlike the two previous experiments, this one **did not** pit a man against an identical female. Instead, it pitted a female who took a one-year parental leave against an identical female who did not take a leave (Conditions 9.1 and 9.2), and a male applicant who took a one-year parental leave against an identical male who did not (Conditions 10.1 and 10.2). This was done to disentangle the effect of gender of applicant from the effect of whether the applicant took a parental leave, an analysis that cannot be accomplished with the cross-sex comparisons used in the prior experiments, because it requires comparing same-sex applicants to avoid confounding two variables (gender and leave status). These four conditions also permitted a direct comparison of either two men or two women who are identical except that one of each pair is described with female gendered adjectives and the other with male adjectives.

Conditions 9.1/9.2 And 10.1/10.2: Evaluating Same-Sex Competing Candidates Who Either Took Versus Did Not Take One-Year Parental Leave

Total N for Conditions 9 and 10 = 204 faculty respondents, 109 females and 95 males.

Total N for Condition 9 = 93 faculty, 46 men, 47 women.

Total N for Condition 10 = 111 faculty, 49 men, 62 women.

9.1 Dr.X=woman described in female adjective condition, 2 children and attorney husband; Dr.Y=male foil; Dr.Z=woman described in male adjective condition, 2 children and doctor husband, 12-month leave.

9.2 Dr.X=woman described in male adjective condition, 2 children and doctor husband; Dr.Y=male foil; Dr.Z=woman described in female adjective condition, 2 children and attorney husband, 12-month leave.

10.1 Dr.X=man described in female adjective condition, 2 children and attorney wife; Dr.Y=female foil; Dr.Z=man described in male adjective condition, 2 children and doctor wife; 12-month leave.

10.2 Dr.X=man described in male adjective condition, 2 children and doctor wife; Dr.Y=female foil; Dr.Z=man described in female adjective condition, 2 children and attorney wife, 12-month leave.

To recap, in 8 of the 10 conditions (16 of the 20 versions) faculty respondents received a contest in which two males and one female competed for a position, and each was depicted with different adjectives (i.e., personae) to further obscure the gendered contrast (which was done in a counterbalanced manner so that the same adjectives were used just as often for male applicants as for female applicants but this was implemented between-subjects so faculty were unaware that peers received the mirror versions). Respondents' comments revealed their lack of awareness of the design and hypothesis; a subset of faculty asked to guess the hypothesis of the study were unable to do so.

c. Cover Letter and Experimental Materials. The materials sent to each faculty respondent consisted of a search-committee chair's narrative summary of the scholarly record, job talk, and interview for the three top hypothetical candidates for a tenure-track assistant professorship (samples of materials appear in the Resources section). The chair's summary described the mean rating given by the members of the department for each of the three hypothetical finalists, based on research publications, job talk, reference letters, and interviews with individual faculty. The chair's summary did not list the number or type of publications, because pilot testing revealed that publication expectations differed dramatically across fields and subfields and had different value across institutions (e.g., what one institution regarded as a desirable number of publications to make their short-list differed from another, with research-intensive ones expecting greater research productivity than teaching-intensive ones). So, the chair's summary provided faculty ratings of 9.5 for the two strongest applicants, meaning they were very highly regarded by departmental faculty who studied their CVs and attended their talk and interview, without specifying their number or type of publications. Faculty subjects were sent individual personal emails from the authors—no web interface, Qualtrics, or other "processed" format was used, so we were able to determine the exact number of faculty who were invited to participate but chose not to. Respondents were instructed to imagine that their own departmental colleagues rated the hypothetical applicants on a 10-point scale and the narrative summary for each candidate was based on their evaluation of the applicants' research prowess, strength of references, job talk ratings by colleagues, and individual interviews. Respondents read that their departmental colleagues rated these three hypothetical finalists for a tenure-track assistant professor position in their department between 9.3-9.5, which was in the excellent range.

d. Ranking Three Short-Listed Candidates. Faculty respondents were asked to rank the three short-listed applicants in order of their hiring preference, from first (top) to third, and also to rate them on a 10-point scale similar to that of their hypothetical colleagues who rated them between 9.3-9.5. They rated their top-ranked applicant very highly on the same 10-point scale on which their hypothetical departmental colleagues had rated them 9.3-9.5. Mean ratings revealed that 90.5% of faculty rated their top-choice applicant between an 8 (excellent) and a 10 (extraordinary), on a scale in which 3=acceptable. We note that not a single respondent mentioned that the materials were unrealistic or insufficient to do their rankings, when invited to comment, thus validating the national consultants' opinions regarding the feasibility of using the narrative summaries for this task. Given the competitive nature of tenure-track positions, the finalists are extremely competent, so the fact that faculty rated these applicants as excellent conforms to this expectation. Many faculty respondents spontaneously commented on how competitive tenure-track searches in their own departments are, reporting they routinely have to reduce an applicant pool of hundreds of talented PhDs to just three when there are often 30+ who are superb. For example, "In the end, only one candidate can be selected, and then there is the overall problem … we have a glut of outstanding, well-trained candidates and there are not enough jobs to go around" (from a senior Chair-holding Professor of Biology).

As noted above, each faculty respondent was asked to rank only one group of three short-listed applicants. The foil, Dr. Y, was pre-rated slightly lower (9.3 versus 9.5 for Drs. X and Z), and was chosen only 2.53% of the time, thereby allowing us to focus on the choice between Dr. X and Dr. Z for 97.5% of the faculty rankings. As already noted elsewhere, respondents' comments indicated a lack of awareness that the identical short-list they rated was sent to other faculty with Dr. X's and Dr. Z's gender and lifestyle reversed in a fully counterbalanced design so that we could detect potential hidden biases.

IV. Statistical Analysis and Related Issues

a. 21-Word Solution. We open with the 21-word solution proposed by Simmons, Nelson & Simonsohn (17): "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study."

b. Raw Tally of Votes (see next page).

Table S5. Raw Tally of Votes by Condition. Overall N=873; 711 in main study (Experiments 1-3), 35 in full-CV validation (Experiment 4), 127 in single-applicant-rating study (Experiment 5); data aggregated across faculty gender and subconditions.

CONDITION (".1"and ".2" versions counterbalanced for sex) (N=873)	**N ranking woman #1**	**N ranking man #1**	**N ranking foil #1**	**N giving tied ranks**	**N missing data**
EXPERIMENT 1: Matching Lifestyles					
1.1/1.2—Single/no kids	44	18	2	1	0
2.1/2.2—Married/no kids	45	17	0	1	1
3.1/3.2—Married/stay-at-home spouse	39	15	0	2	0
4.1/4.2—Married/home-business spouse	27	24	1	4	0
5.1/5.2—Married/spouse working outside home	35	17	5	3	0
6.1/6.2—Divorced/with kids	38	20	1	2	1
Total Conditions 1-6=363	228	111	9	13	2
EXPERIMENT 2: Contrasting Lifestyles					
7.1/7.2--Married Father vs. Divorced Mother **Total Condition 7=69**	35	28	2	4	0
8.1/8.2--Married Father vs. Single Childless Woman **Total Condition 8=75**	52	17	2	3	1
Total Conditions 7-8=144	87	45	4	7	1
EXPERIMENT 3: Contrasting Same-Sex Candidates With vs. Without 1-Year Parental Leave	**N ranking candidate withOUT leave #1**	**N ranking candidate WITH leave #1**	**N ranking foil #1**	**N giving tied ranks**	**N missing data**
9.1/9.2 (X & Z both women) **Total Condition 9=93**	43	46	1	2	1
10.1/10.2 (X & Z both men) **Total Condition 10=111**	49	52	4	3	3
Total Conditions 9-10=204	92	98	5	5	4
EXPERIMENT 4: Full-CV Validation	**N ranking woman #1**	**N ranking man #1**	**N ranking foil #1**	**N giving tied ranks**	**N missing data**
Total Condition 11=35	25	8	2	0	0
EXPERIMENT 5: Single-Applicant Rating on 1-to-10 Scale	**Mean rating--female**	**Mean rating--male**	**NA**	**NA**	**N missing data**
Total Condition 12=127	8.20	7.14	NA	NA	0

c. Logistic vs. OLS in the Interpretation of Interactive Effects. Data from the 711 respondents in the first three experiments were analyzed in a series of logistic regression models using a SAS Genmod procedure, to model the probability that a given female applicant was preferred over her male competitor—both with and without controls and including interaction terms. (Note that Experiment 3 conditions 9.1/9.2 and 10.1/10.2 examined the "took leave" variable as the target of interest instead of the "female gender" variable.) Additionally, OLS models were run to aid the interpretation of regression coefficients in interactions (18-20) and to exclude the possibility of biased standard errors on interaction terms in nonlinear models. Finally, separate weighted logistic regressions were run to test for nonrandomness.

In the main text we reported the results from the unweighted logistical models, because they are more intuitively translatable into odds-ratios of being hired and avoid the suspicion some readers may harbor when data are transformed. Such suspicion might occur because there are far fewer women faculty in two of the fields—engineering and economics; thus the probability of sampling any given woman is higher than the probability of sampling a corresponding male. This results in increasing the males' weights over females', from 50-50 in the unweighted analysis to roughly 70-30 in the weighted models.

Here we report the weighted analyses, with the weights described below. Power analyses of the optimal sample size needed to detect large effects (based on the published literature) involving the focal hypotheses of gender of applicant x gender of faculty, without biasing results toward obtaining significance, were supplemented by post hoc power considerations that combined the samples for three of the five experiments (excluding Experiment 3 which involved same-sexed contrasts and Experiment 5 which involved rating a single candidate). As noted earlier, although overall response rates in the 30-35% range are typical in experimental surveys (7, 8), to check on the representativeness of the 34% sample of respondents we conducted two types of validity checks and together they suggest that the 34% of respondents reflected the hiring preferences of the entire sample. The weighted analysis further strengthens this claim.

As planned, the foil (Dr. Y) was chosen by only 2.53% of respondents, so all analyses are based on the 97.5% of respondents who preferred one of the two slightly stronger applicants, Dr. X or Dr. Z. Depending on the specific hypothesis being tested, variables used in the logistic models included: gender of faculty rater, field of faculty rater (biology, economics, engineering, psychology), gender of hypothetical applicant, lifestyle of hypothetical applicant (marital status, children, parental leave, etc.), and gendered personae of hypothetical applicant (stereotypical male persona vs. female persona adjectives crossed with hypothetical applicant gender). By way of simplifying these analyses, each of the conditions was analyzed separately, as well as analyzed while pooled across the six conditions of Experiment 1.

As noted above, the final two conditions of Experiment 3 (each with two subconditions) pitted same-sex applicants against each other (leave-takers vs. non-leave-takers of the same sex) while the other eight conditions (16 versions) of the first two experiments analyzed cross-sex competitions (e.g., single female vs. single male; married father with stay-at-home spouse vs. married mother with stay-at-home spouse). The models for these experiments included all two-way interaction terms: lifestyle x rater gender, lifestyle x rater field, rater gender x rater field (3-way interaction is not estimable in the full model.) The various models were based on Ns ranging from 69 to 363 depending on the experimental conditions reflected in the evaluation of a given hypothesis (note that a small number, usually 2-3%, of subjects dropped out of individual analyses due to missing data or tied rankings, thus columns do not always sum perfectly). The anonymized data are archived on the webpage of the Cornell Institute for Women in Science (www.ciws.cornell.edu) where they are available for downloading and reanalysis.

d. Weighted-Analysis Results. In nearly all respects, the weighted analyses corroborated the unweighted findings reported in the main text. These analyses are also archived (www.ciws.cornell.edu). As will be shown, these analyses revealed an overall strong preference for hiring female applicants over identically-qualified males who shared the same lifestyle, and the effect sizes remained large. Type 1 and Type 3 logistic regression models were constructed for each analysis; all interactions, when reliable, were further tested with Bonferroni correction. Overall, across the six lifestyles, collapsing across rater gender, rater field, candidate lifestyle, and gender of persona used to describe the candidate, the weighted analysis revealed that women applicants were strongly preferred by roughly a 2-to-1 ratio. Although weighted analyses can be performed in a variety of equally-defensible ways, an independent statistician requested these raw data through a third party

associated with the peer review process in order to replicate the results. His analyses did in fact replicate these findings using R rather than the SAS we used. His R files and results are also archived for interested readers.

Specifically, comparing female and male applicants who were identical in both scholarly quality and lifestyle (i.e., collapsing across Conditions 1.1/1.2-6.1/6.2), women were ranked #1 in the weighted analysis by 65.88% of faculty respondents, N=337, chi-square=34.90, $p<0.0001$. (In the unweighted analysis reported in the main text this female preference was 67.27%, chi-square 40.38, $p<0.0001$). Women were preferred more often than identical men in biology (69.17%, chi-square 15.11 vs. 69.23% in the unweighted analysis, chi-square 13.46, $p<0.0002$), engineering (63.89%, chi-square 7.88 vs. 66.67% in the unweighted analysis, chi-square 9.33, $p<0.002$), and psychology (72.20%, chi-square 17.56 vs. 72.83% in the unweighted analysis, chi-square 19.17, $p<0.001$). As was found in the unweighted analysis, the sole exception was the field of economics, in which only the female faculty rated identically-qualified female candidates #1 significantly more often than they rated identical males (69.05% vs. 30.95%), which was similar to the 68.3% female preference reported for female economists in the unweighted analysis. Once again, only male economists did not differ in their ranking of identical male and female candidates (53.97% vs. 46.03%, respectively, n.s.), which was similar to the 54.84% vs. 45.16% reported in the unweighted analysis. Whereas this sex difference in female-candidate preference between male and female economics-faculty-raters was shy of reaching significance in the weighted analysis ($p=0.089$), it did reach significance in the **un**weighted analysis: chi-square=3.89, $p<0.049$. Thus, overall these weighted analyses closely mirrored the unweighted results reported in the main text.

The most common lifestyle for assistant-professor applicants is single without children. As was true in the unweighted analysis, in the weighted analysis single women were strongly and equally preferred by male and female faculty by ratios between 3-to-1 and 4-to-1. In the main text we reported the results of a contest between a divorced mother with two preschool-age children pitted against a married father of two whose spouse is a stay-at-home mother. Another contrast was of a married father with stay-at-home spouse competing against a single woman with no children. In the weighted analyses it was again found that male faculty raters preferred the married father over the identically-qualified divorced mother (59.50% vs. 40.50%, respectively; note that the figures for weighted versus unweighted differ at the second decimal place only—unweighted is 59.46 vs. 40.54), whereas female faculty raters preferred the divorced mother over the identically-qualified married father (73.02% vs. 26.98%, respectively; N=63, chi-square=5.412, $p=0.02$). This closely mirrored the results from the unweighted analysis (chi-square=6.12, $p<0.012$). When a married father was pitted against a single woman without children, everyone preferred the single woman—female raters preferred her 71.07% of the time and male raters preferred her 68.12% of the time (N=69, chi-square=10.55, $p=.0012$). This finding also corroborated that found in the unweighted model, chi-square=15.34, $p<0.0001$.

Concerning the effect of taking a one-year parental leave during graduate school, the results of the weighted analysis mimicked the unweighted results. Holding constant applicant quality, but pitting an applicant who took a one-year parental leave against a *same-sex* candidate who also had children but took no extended leave, once again showed that male and female faculty raters responded differently to hypothetical candidates based on candidate gender and leave status, though marginally (N=190, chi-square=3.65, $p=.056$ for the weighted analysis, vs. chi-square=4.2, $p<0.04$ in the unweighted analysis). Male faculty raters preferred mothers who took extended leaves 2-to-1 over matched-academic-quality mothers who took no extended leaves (68.8% to 31.2%, $p<0.01$), but these male faculty raters showed no preference between fathers who took vs. did not take extended leaves (48.9% vs. 51.1%, $p>0.10$). Female faculty raters also showed no preference regarding fathers' leave status (55.3% with leave vs. 44.7% with no leave, similar to what was found in the unweighted analysis--53.6% vs. 46.4%). Also, in the weighted models, female raters <u>marginally</u> preferred mothers who did not take a leave over those who did by a margin of 62.7% to 37.3%, closely shadowing the findings in the unweighted analysis, 62.2% to 37.8% for mothers who took <u>no</u> extended leave over mothers who took leave—although the error bars slightly overlapped for this contrast so it is best to view it as suggestive rather than confirmed. In sum, in a contest between mothers with identical qualifications, male and female faculty have different, with the interaction of gender of rater x gender of applicant hovering around $p<0.05$.

The bottom line is that the female preference reported for the unweighted analysis in situations in which men and women applicants were identical and shared the same lifestyle was extended to situations in which the applicants occupied different lifestyles, with both sets of comparable findings confirmed in the

weighted results. When a married man with two kids and a traditional stay-at-home spouse was pitted against a single woman with no kids, faculty strongly preferred the single woman (68.95% in the weighted analysis, chi-square=10.55, $p<0.012$) vs. 73.24% in the unweighted analysis reported in the main text). The rather counter-intuitive finding in the main text showing that female faculty preferred a divorced woman with two preschool-aged children over a married man with two kids and a traditional stay-at-home spouse was also found in the weighted analysis. When broken down by the gender of the faculty respondent, female faculty strongly preferred the divorced mother (73.02% in the weighted results vs. 71.43% in the unweighted analysis, both *p's*$\leq$0.02) whereas male faculty preferred the married father in both analyses (59.50% in weighted and 59.46% in unweighted, respectively, both *p's*<0.05).

e. Results by Field. Collapsing across rater gender, women had a 69.23% chance of being preferred for hiring in biology (chi-square=13.46, $p=0.0002$), a 72.53% chance in psychology (chi-square=18.47, $p<0.0001$), a 66.67% chance in engineering (chi-square= 9.33, $p=0.0023$), and a 58.57% chance in economics (chi-square=2.05, $p=0.15$)--although female economists also significantly preferred women candidates (at 67.5%) and they were not reliably different from men and women faculty from the other fields in their preference for hiring them (all *ps* >0.10). Summing across all fields, the overall 2-to-1 female preference (67.26% for women candidates vs. 32.74% for men candidates) was highly reliable (chi-square=40.38, $p<0.0001$), and was also found for a comparison of identical applicants within five of the six different lifestyles, each of which showed the same strong female preference (e.g., being married or single, with or without children did not change the 2-to-1 female advantage, N=339, chi-square (5)=6.41, $p=0.269$).

f. Results by Lifestyle. In the overall model, the lifestyles depicted in Conditions 1.1/1.2 through 6.1/6.2 showed equivalently strong hiring preferences for women, notwithstanding the aberration for male economists: Summing across the six lifestyles, there were no effects beyond the strong main effect for preferring female applicants. While it is possible to decompose a given cell of the design and test for aberration from the overall finding, when such interactions do not reach significance in the overall model they are not decomposed further. Given that while not controlling for any other variables, there is no evidence of differential hiring of women according to lifestyles in the overall model—i.e., no lifestyle differed significantly from the others--and lifestyles did not differ as a function of either rater gender and/or rater field, it is unwise to overinterpret the rare aberration, such as that observed when raters of both genders appeared to favor a male applicant whose wife worked in home-based businesses, representing one cell out of 12 relevant cells. In sum, while controlling for both rater gender and rater field, Experiment 1 did not show evidence of differential hiring of women according to lifestyle. The model testing this entailed the following two-way interaction terms: lifestyle x rater gender, lifestyle x rater field, and rater gender x rater field. (The 3-way interaction was not estimable in the full model.)

The clearest test of sex differences in hiring preferences comes from the contrast between male and female applicants identical in quality who share a lifestyle, as just described; this "gender only" manipulation is how this question has been studied in prior experimental research (for an early review of this literature see 19). The most common lifestyle for applicants for assistant professorship is single without children; in this lifestyle, women are strongly and equally preferred by male and female faculty, 66.7% and 75.9%, respectively ($p=.43$). However, in the real world, competing candidates do not always share the same lifestyle. Sometimes applicants are married, married with children, divorced with children, and have partners who work inside versus outside the home. Thus, in the two experimental conditions of Experiment 2 we explored lifestyle contrasts of candidates that are likely to occur in the real world to see if there are situations in which female applicants are downgraded. In one of these cross-lifestyle contrasts the 2:1 female preference changed.

One contrast that did not result in a consistently strong female preference is of a divorced mother with two daycare-age children whose ex-husband does not plan to relocate with her for her job, pitted against an identically-qualified married father of two children whose spouse is a stay-at-home mother. Another such contrast is of a married father with stay-at-home spouse competing against an identically-qualified single woman with no children. We investigated how candidates identical in quality but differing along these lifestyle dimensions varied in hirability. We found that male faculty raters preferred the married father over the identically-qualified divorced mother (57.1% vs. 42.9%, respectively), whereas female faculty raters preferred the divorced mother over the identically-qualified married father (71.4% vs. 28.6%, respectively; N=63, chi-square=6.12, $p=0.013$). Given that in math-intensive fields there are many more male faculty, this suggests a

barrier for divorced mothers of preschoolers applying for positions. (Additionally, male faculty raters downgraded divorced fathers vis-à-vis identically-qualified divorced mothers, 39.3% preference for divorced fathers vs. 60.7% preference for divorced mothers, chi-square=5.14, p=.02. Women raters also preferred the divorced mother 70.0% of the time, which was equivalent to men's preference for the divorced mother). As seen in Figure 3, in the competition between a married father and single woman without children, however, everyone preferred the single woman by just over a 2:1 ratio for male raters and a 3:1 ratio for female raters (71.0% by male faculty; 75.8% by female faculty; N=69, chi-square=.8, p=.54).

g. Analysis of Effect of One-Year Parental Leave. One lifestyle factor of current national policy interest is the effect on applicants of either gender of taking a one-year parental leave during graduate school. Holding constant applicant quality, but pitting an applicant who took a one-year parental leave against a same-sex candidate who also had children but took no extended leave, revealed that male and female faculty raters responded differently to applicants based on applicant gender and leave status (N=190, chi-square=4.2, p<0.05). Male faculty raters preferred 2-to-1 mothers who took extended leaves over matched-academic-quality mothers who took no extended leaves (65.9% to 34.1%), but these male faculty raters showed no significant difference in preference between fathers who took vs. did not take extended leaves (48.9% vs. 51.1%, p>.10). Similarly, female faculty raters also showed no significant difference in preference between fathers' leave status (preferring 53.6% fathers with leave vs. 46.4% fathers with no leave, p>.10). However, female raters showed a marginal leave-based preference in evaluating female candidates, preferring women candidates who took no extended leave (62.2% to 37.8%), which was the flip side of male raters' 34.1% to 65.9% preference for non-leave taking mothers (N=190, chi-square = 7.05, p<.01). Thus, in a contest between mothers with identical qualifications, male faculty prefer those who take extended leaves, but female faculty do not.

h. Carnegie Classification Findings. There was no effect for Carnegie classification: all three types of institutions exhibited comparable preference for hiring females over identical males, and this did not interact with field or faculty gender. This is interesting because one might expect that small teaching-intensive colleges would differ from large research-intensive institutions in their female preference, but this was not the case. Thus, in fields that are math-intensive as well as those that are not, and in fields in which women are already well-represented as well as ones that are not, faculty exhibited a strong preference for female applicants to tenure-track posts over their identical male counterparts, and this was true in all three types of institutions.

V. Interpretative Issues: How Our Findings Compare to Past Research

a. Do Staff Lab Managers Become Tenure-Track Professors? The 2-to-1 pro-female preference reported in this study, which was fairly consistent across gender of faculty raters, fields, lifestyles (with a few exceptions), and types of Carnegie institution, runs counter to studies reporting a strong anti-female bias in hiring. Two of these anti-female bias studies deserve mention. First is an experiment conducted with 127 science faculty at six U.S. universities, who were asked to evaluate hypothetical applicants with "ambiguous" credentials who were finishing undergraduate degrees and applying for a full-time staff lab manager post (8). In this study faculty rated male applicants higher and recommended higher starting salaries and more mentoring than they did for female applicants--even though there was no difference between their applications. Both female and male faculty raters exhibited this bias. Numerous other experiments have also reported bias against females' teaching skills and work products (e.g., 21, 22, 23).

The above experiments and many others deal with biases against female undergraduates, e.g., undergraduate applicants with ambiguous academic records who apply for lab-manager posts (just described) or summer jobs (21), or undergraduates rating work products or teaching effectiveness of lecturers who are comparable except for gender (e.g., 22). While these findings are interesting and important, for several reasons they seem unlikely to generalize to the hiring of tenure-track professors.

In contrast to ratings of students for fairly short-horizon positions or work products, tenure-track hiring of prospective faculty entails decision-making for long-term investments by current faculty members. The reason this distinction matters is because finalists for tenure-track positions are accomplished scholars; they have already demonstrated success in completing doctoral programs and accruing publications and strong letters of support. As noted earlier in the Supplement, numerous faculty respondents in these experiments spontaneously commented on how competitive tenure-track job searches in their departments are, with hundreds of talented applicants vying for a single position. Contrast this with an applicant for a staff lab-manager post who

was depicted as "ambiguous", with an academic record that was equivocal (8)--unlike doctoral candidates for tenure-track positions who have already demonstrated success. The reason these researchers depicted the lab manager applicant ambiguously is precisely because they wanted to maximize the chance of detecting anti-female bias in a situation likely to arouse it, reasoning that bias was most likely to occur in ambiguous situations as opposed to situations in which candidates were competent and strong (such as the situation when candidates for a tenure track short list are compiled). This can be seen in these authors' words:

"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." (p. 1 Supporting Online Materials at http://www.pnas.org/content/suppl/2012/09/16/1211286109.DCSupplemental/pnas.201211286SI.pdf#nameddest=STXT).

Thus, one possible reason the present study did not find anti-female bias among faculty raters of tenure-track applicants is because the applicants for the tenure track positions are, in fact, unambiguously excellent: the three short-listed applicants were described as possessing the sort of competence that short-listed candidates for tenure-track positions usually possess according to faculty respondents—i.e., they successfully completed doctoral training, published research, garnered supportive letters, and were described in the chair's summary as 9.5 on a 10-point scale. Perhaps if we had depicted Drs. X, Y, and Z as, say, 3s on the 10-point scale (acceptable but not irrefutably excellent) instead of 9.3-9.5 (impressive), this would have changed our results. But this would also have been unrealistic, since such weakly-rated candidates would not typically be finalists in the academic job market containing many highly-qualified women and men seeking tenure-track positions.

If bias does exist when evaluating ambiguously-described candidates with unremarkable records, perhaps this could ultimately have deleterious consequences for shaping the academy by culling marginal female undergraduates out of the graduate school pipeline. However, few scientists in the academy report that they preceded their graduate careers by working as full-time staff lab managers. (We polled 83 academic scientists randomly sampled from the same institutions as used in our study, and asked them if they ever worked as full-time staff lab managers before beginning their graduate careers, only 1 stated that she did. Most of the others noted that they went directly from undergraduate to graduate school or, if they ever managed or directed a lab, it was as a full-time graduate student who had already been accepted into graduate school on the basis of her/his excellent record, and not as a full-time staff employee.) This does not minimize the importance of anti-female bias where it exists, but it does caution us not to assume it is responsible for deleterious effect in shaping the academy's demographics, given that the vast majority of scientists in the academy did not begin as staff lab managers, nor were they ambiguously-competent academically as undergraduates. Rather, evidence for anti-female bias in hiring lab managers should be countered by initiatives directed at such hiring and not generalized to account for the underrepresentation of women among tenure-track faculty. As we note below, there is no evidence internal to this study or in the actual real-world hiring data that supports the claim that tenure-track hiring is anti-woman.

There is one instance of experimental evidence of bias in the case of faculty ratings of female applicants for a tenure-track faculty position (23). This experiment used actual faculty as raters, as was done in the present study. It was published 16 years ago and it was confined to one field, psychology. Psychology is not one of the fields of concern in the underrepresentation debate because it does not have low numbers of women; In 2010 women comprised 66% of tenure-track assistant professorships in psychology and were 71% of Ph.D. recipients (see 2, Figure 3b). It is possible that gender equity and the desire to diversify faculty is greatest in fields in which women are the most underrepresented, unlike psychology. However, we included the field of psychology in our experiment in order to see if we could replicate this 16-year-old finding, using a nationally representative sampling plan and procedures. Yet, even among the large group of psychologists in our sample we detected no anti-female bias. In fact, our results did not differ as a function of the level of female representation in a given field. Interestingly in the context of the above discussion regarding candidates who are unam-

biguously competent, in this 16-year-old study there was a condition in which faculty evaluated male and female applicants for early tenure--who were clearly unambiguously competent. They found no sex bias in rating this group.

In sum, the present study is the largest of its kind, and the only one to examine tenure-track hiring in math-intensive fields, utilizing a large, national sampling plan (including faculty respondents from all 50 states and the District of Columbia) and validity checks on response bias. And in contrast to some claims, it found no evidence that tenure-track faculty hiring was biased against women applicants. In fact, it found evidence that women fared better than identically-qualified men.

b. Do These Results Differ from Analyses of Actual Hiring Data? In view of the fairly consistent finding in the present study of an approximately 2-to-1 ratio for female-to-male hiring preference, one can ask how this result stacks up against other data, especially data from real-world hiring practices. The answer is that it stacks up fairly well. There are a number of large-scale (national) analyses of hiring that, as we describe below, show a female preference, and there are a number of local analyses (based on multi-year hiring for one or two universities) that also show that female applicants are hired at higher rates than then their male competitors. Most of these analyses show female preferences smaller than the 2-to-1 preference observed in these experiments but several also show preferences that large. Importantly, none of these real-world hiring analyses controlled candidate quality and lifestyle or any of the other variables that were controlled in the present experiments. Despite this, these large-scale hiring analyses consistently show that even though women are less likely to apply for tenure-track positions, when they do apply they are hired at higher rates than men.

A number of audits of hiring by universities have been reported in the past two decades and these have reported either a neutral playing field in non-mathematical fields in the sense that the status predictors worked equivalently across genders (e.g., 24, p. 786), or, more commonly, a pronounced female hiring advantage in math-intensive fields. Before describing these data, however, it is important to make the point that these studies are open to alternative interpretations. These data were not collected under controlled experimental conditions and therefore the authors were not able to rule out all of the factors that could have a causal role in female preference, such as applicant quality (at best proxied by the total number of publications or type of journals--and usually not even by these) as opposed to the perfect quality control possible in experimental studies which employ identical applicant dossiers. These hiring analyses also do not control for the number of job offers an individual woman or man receives--as opposed to the number of acceptances, which is what is measured in these real-world hiring analyses. This potentially underestimates how many faculties voted to OFFER their top candidate of a given gender a job, because one can only accept one of possibly multiple offers, a point elaborated upon below. Indeed, these real-world confounds were a driving force behind the present experimental study--to address precisely these uncontrolled factors.

Here is what we know about the female advantage in real-world hiring of tenure-track applicants in STEM fields in the United States and Canada: There is a female advantage in all large-scale studies dating back to the 1980s, but with a few exceptions the female advantage is less than the 2-to-1 finding reported here. Consider: An NRC national survey of six math-intensive disciplines examined faculty experiences and institutional policies in place during 1995-2003 (25). It included over 1,800 faculty members' experiences in nearly 500 departments at 89 research-intensive (formerly called R1) universities in the U.S. Although a smaller proportion of female Ph.D.s applied for tenure-track positions at these R1 universities, those who did apply were invited to interview and offered positions more often than predicted by their fraction of the applicant pool. For example, in the field of Mathematics, only 20% of applicants for tenure-track positions were females, but 28% of those invited to interview were female, and 30% of those offered tenure-track positions were females. In all six math-intensive disciplines studied by the NRC, female applicants were invited to interview and offered positions at rates higher than those for men. But the ratios are lower than the 2 to 1 level found in the present experimental study as can be seen in Table 1 which is based on the NRC data.

FIELD	Mean % Female Applicants	Mean % Invited to Interview	Mean % Offered Position
Physics	12%	19%	20%
Biology	26%	28%	34%
Chemistry	18%	25%	29%
Civil Eng.	16%	30%	32%
Electr. Eng.	11%	19%	32%
Mathematics	20%	28%	32%

Table S6. Fraction of female applicants for tenure-track positions invited to interview and offered positions at 89 U.S. research universities (NRC, 2009, p. 8, adapted from Findings 3-10, 3-13).

Another large-scale analysis of real-world hiring was based on NSF's Survey of Doctorate Recipients (SDR). The SDR is the largest, continually replenished panel data on the trajectories of U.S. scientists. To determine the likelihood of getting a tenure-track job, researchers analyzed over 30,000 respondents interviewed between 1981-1995 (26). Similar to the NRC findings above, women were less likely to apply for tenure-track positions. However, their low number was not due to gender per se, but rather family variables associated with it. Controlling for such variables--such as presence of young children--revealed that women during this epoch were also hired at rates comparable to or higher than those for men. For example, in the largest demographic, unmarried women without children were 16% more likely to get tenure-track jobs than were unmarried men without children (26).

An analysis of all tenure-track hiring over a six-year period in 19 STEM fields at a large U. S. state university also found that female applicants were far less likely to apply for these positions (27). However, when women did apply, they were just over 2-to-1 times more likely to be offered jobs as men: of 3,245 applicants for tenure-track positions in 19 STEM fields between 2000 and 2006, 2.03% of male applicants were hired versus 4.28% of females (27). In a similar analysis of academic hiring at a large Canadian university between 1991 and 1999, a similar conclusion was reached, with women being just under 2-to-1 times more likely to be hired: "Over the 8 years, on average: 5.4% of female applicants were appointed compared to 2.9% of male applicants; 21.7% of female applicants were interviewed compared to 15% of male applicants; and 24.9% of female applicants who were interviewed were hired whereas 19.2% of men who were interviewed were appointed. Again, the results in each of the years are remarkably consistent. Women had almost twice the chance of being hired as did men" (27).

An analysis of academic hiring throughout Canada over the period between the late 1960s and early 1990s was carried out using national Statistics Canada data. It reported that the percentage of female job recipients was typically higher than their fraction of the job applicant pool (28). For example, between the early 1980s and the early 1990s Irvine estimates that women constituted 25.6% of the academic hiring pool but were actually hired for 33.4% of the assistant professorships, leading him to conclude that: "Another analysis of hiring data from 36 departments at two large Canadian universities replicated the above results: in the reports about three most recent positions filled in each unit, the percentage of female applicants was 29% but the percentage of women actually hired was 41%" (28). In sum, actual large-scale hiring analyses all show either a level playing field or, more commonly, a female advantage (see also 29, 30), but usually at less than the 2-to-1 ratio reported in the present experimental study. Below we consider three reasons for this difference between the present experimental results and actual hiring analyses.

Finally, in an analysis of all academic hiring at U.S. law schools between 1986-1991, Merritt and Reskin (31) determined that of the 1,094 positions for which someone was hired, in their analysis women were less likely to be hired at more senior titles (associate and full professor) and also less likely to have the most prestigious course assignment (constitutional law) as opposed to less prestigious courses (e.g., family law, appellate advocacy, trusts and estates). However, when it came to getting a tenure-track job, women fared as well as or better than male counterparts, obtaining more jobs at the most prestigious law schools even after controlling for background variables such as prestige of JD institution, experience as a federal court of appeals or United States Supreme Court clerk, age, membership on law review, etc.:

"When we analyzed institutional prestige as an ordinal variable, white women and men of color obtained jobs at significantly more prestigious institutions than did white men with comparable credentials. This result suggests that law schools used affirmative action programs to grant modest hiring advantages to white women and men of color, preferring them over equally qualified white men (p. 274)…White women also received some preference in hiring, although the advantage was smaller than that afforded minority men. Our analysis suggests that if a white man and white woman entered the academic market with identical and exceptional credentials, the woman might win a job at Harvard (tied in prestige for third at 3.76) while the man would secure an appointment at Berkeley (ranked eighth at 3.37). A white woman with somewhat weaker credentials might teach at the University of Colorado (rated 2.27) while a man with those same credentials would begin teaching at Fordham University or the University of Georgia (both rated 1.82)" (31, p. 151).

***c. Uncontrolled Applicant Quality*.** One potential cause of the disparity between the present 2-to-1 experimental results and actual large-scale hiring analyses that typically find a somewhat smaller female advantage has to do with applicant quality. This has not been satisfactorily controlled in any of the actual hiring analyses, but in the present experiment it was perfectly controlled because identical accomplishments were used for both genders in a between-group design. On this basis, one could possibly argue that in the actual real-world large-scale hiring analyses the average woman is not as high in quality as the average man who competes against her for a tenure-track post. If true, then any inherent desire on the part of faculty to enhance gender diversity might nevertheless result in a lower number of females hired due to their lower quality than would be found in an experiment in which the quality of male and female applicants was identical. We are not arguing that this is the case, but merely pointing out that on logical ground it is a possible factor that could be responsible for the generally higher female advantage when identical scholarly records are used for males and females than in the actual real-world hiring where there is little or no control on quality. Others have made such an argument many years ago but today there are grounds for thinking that the quality of male and female applicants is comparable (28). Regardless of the validity of the claim, it is unsatisfactory to use undifferentiated proxies for applicant quality such as total number of publications or impact of the journals (or even number of publications adjusted for the applicant's position in the list of authors). Even two male applicants who have the same number of publications in journals with the same impact factors can differ in their hirability because of the area of research, the methodology, the judged importance of the problem, etc. And yet, even this imperfect type of statistical control is rare in non-experimental analyses, which usually are unable to control quality at all.

***d. Real-world hiring analysis is based on job acceptance rates, not rates of job offers*.** Actual hiring analyses may underestimate the strength of female preference in another way. Such analyses use as the dependent variable the number of men and women who were hired--not the number of job offers to them. If there is a female preference than it is likely that the men and women who get offers may get multiple offers, and because men on average receive fewer offers, women's job acceptance rates may underestimate their number of actual offers. Because individuals can each accept only one offer, this fact could obscure the strength of a female preference. If multiple faculties voted to offer their top woman candidate the job over a man—she can only accept one job offer, and this is only counted as one woman hired in the national hiring dataset. There is some suggestion that something like this was operating in the NRC analysis. Once again, we are not arguing that this is the case, but it is simply a logical possibility that, if true, might result in differing estimates between experimental and observational studies.

In sum, the strong (2-to-1) female preference under controlled conditions appears to be a robust finding across nearly all of the conditions of these experiments, and it accords fairly well with the actual hiring data, which although usually displaying less than a 2-to-1 preference for women, sometimes reaches this level and, importantly, never shows a preference for men.

e. Causes of Underrepresentation. One might wonder if the claim of sex bias in hiring is a "straw man", as one of the consultants on this study opined. Does anyone believe that women are discriminated against in academic hiring and that this is the cause of their dearth? And if so, how can we reconcile such a belief with claims by search committee members that they exercise due diligence (and often more) when it comes to giving serious attention to female and underrepresented minority applicants?

To take the first question, there is a frequently-expressed belief that women are the victims of hiring bias, not just in the past when there may be evidence of bias as some have noted (28), but also in the present,

when evidence is lacking. Numerous commentators have asserted that women face an uphill battle in the academy, including, but not limited to, being invited to interview and hired. Although other barriers are invoked, such as chilly climate, delayed promotion, difficulty balancing work-family, and inequitable salaries, commentators often view the combination of lower numbers of applications from female candidates coupled with anti-woman hiring bias as an important part of the reason for the underrepresentation of women in the academy. Consider:

"Research has pointed to (sex) bias in peer review and hiring. For example, a female postdoctoral applicant had to…publish at least three more papers in a prestigious science journal or an additional 20 papers in lesser-known specialty journals to be judged as productive as a male applicant...The systematic underrating of female applicants could help explain the lower success rate of female scientists in achieving high academic ranks" (American Association of University Women: Hill, Corbett, & Rose, 2010, p. 24).

"It is well-established that the presence of a male or female name on a CV has a strong effect on how that CV is evaluated… In Steinpreis et. al.'s US study, 238 academic psychologists (118 male, 120 female) evaluated a curriculum vitae randomly assigned a male or a female name. Both male and female participants … were more likely to hire the male than the female applicant" (Jennifer Saul, 2012).

"It is now recognized that (sex) biases function at many levels within science including funding allocation, employment, publication, and general research directions" (Lortie et al., 2007, p. 1247).

"These experimental findings suggest that, contrary to some assertions, gender discrimination in science is not a myth. Specifically, when presented with identical applicants who differed only by their gender, science faculty members evaluated the male student as superior, were more likely to hire him, paid him more money, and offered him more career mentoring" (Moss-Racusin, C. Commentary and Analysis from SPSP.org September 21, 2012 http://spsptalks.wordpress.com/2012/09/21/arescience-faculty-biased).

Haslanger provides a table with percentages of women among the faculty of the top 20 graduate programs in philosophy in the U.S., ranging from 4% to 36%, and she concludes that "the data mostly speak for themselves" (2008, p. 214)….and claims that when she was a graduate student there was "a lot of outright discrimination" and that "blatant discrimination has not disappeared" (p. 211). Haslanger, S. (2008) 'Changing the Ideology and Culture of Philosophy: Not by Reason (Alone)', *Hypatia* 23:2, 210-223.

"Psychological research has shown that most people--even those who explicitly and sincerely avow egalitarian views--hold what have been described as implicit biases … There are countless situations in which such mechanisms are triggered: classroom situations, hiring committees, refereeing of papers for journals, distribution of departmental tasks (research, teaching, admin.) etc." Oct. 2, 2010 at http://www.newappsblog.com/2010/10/implicit-biases-1.html

"We are not proposing that gender-blind searches are the only answer (to hiring more women and URMs). We see these as one piece of a larger effort that also involves bias-avoidance training, gender-blind reviews, salary equity adjustments, and a clear examination of bias in the promotion of female professionals." (13, p. 612-613)

"Women and minorities must both deal with implicit bias, a problem that is well documented in the social science literature ... Donna Dean (President of the Association for Women in Science) describes the problem of implicit bias in these terms: 'People are most comfortable with people who think and look like themselves.'" Powell, K. (2007). Beyond the glass ceiling, *Nature, 448*, p. 99.

Thus, the claim of sex bias in the evaluation of females has been prominent in the national debate over women's underrepresentation in some STEM fields. This is often coupled with the finding that a smaller fraction of women than men decide to apply for tenure-track positions, as we noted in the Conclusion to the main article. Countless universities around the U.S. have invested resources to create gender-fair recruitment, hiring, and training (for examples see Resources section at: http://www.ciws.cornell.edu).

REFERENCES

1. Davison HK, Burke MJ (2000) Sex discrimination in simulated employment contexts: A meta-analytic investigation. *J Voc Behav* 56:225-248.

2. Ceci SJ, Ginther DK, Kahn S, Williams WM (2014) Women in academic science: A changing landscape. *Psychol Sci Publ Interest* 1-67. DOI:10.1177/1529100614541236

3. Cuddy, A, Glick P, Fiske ST (2004) When professional become mothers, warmth doesn't cut the ice. *J Soc Issues* 60:701-718.

4. Diekman AB, Eagley A (2000) Stereotypes as dynamic constructs: Women and men of the past, present, and future. *PSPB* 26:1171-1188.
5. Correll SJ, Benard S, Paik I (2007) Getting a job: Is there a motherhood penalty? *Amer Journ of Sociol* 112:1297-1338.
6. Clarkberg M, Moen P (2001) Understanding the time-squeeze: Married couples preferred and actual work-hour strategies. *Amer Behav Sci* 44:1115-1136.
7. Ceci SJ, Williams WM, Mueller-Johnson K (2006) An experimental study of faculty beliefs about tenure, promotion, and academic freedom. *Behav Brain Sci* 29:553-582.
8. Moss-Racusin C, Dovidio J, Brescoll V, Graham M, Handelsman J (2012) Science faculty's subtle gender biases favor male students. *Proc Natl Acad Sci* 109:16474-16479.
9. Holbrook AL, Krosnick JA, Pfent A (2007) in *Advances in Telephone Survey Methodology*, eds. Lepkowski JM et al. (Wiley & Sons, Hoboken), 550.
10. Brown LD et al. (1999) Statistical controversies in Census 2000. *Jurimetrics J* 39:347-375.
11. Schimmack, U. (2012) The ironic effect of significant results on the credibility of multiple study articles. *Psychological Methods* 17(4): 551.
12. Valla JM (2010) Getting hired: Publications, postdocs, and the path to professorship. *APS Obs* 23:10-15.
13. Jones CS, Urban MC (2013) Promise and pitfalls of a gender-blind faculty search. *BioOne* 63:611-612. http://www.bioone.org/doi/full/10.1525
14. Plant EA, Devine PG (1998) Internal and external motivation to respond without prejudice. *J of Personality and Soc Psychol* 75:811-832.
15. Ajzen I, Fishbein M (1980) *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs NJ: Prentice-Hall.
16. Fishbein M, Ajzen I (1975) *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Reading MA: Addison-Wesley.
17. 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. doi:10.1177/0956797611417632 45
18. Karaca-Mandic P, Norton E, Dowd B (2012) Interaction terms in nonlinear models. *Health Services Research*. http://www.thefreelibrary.com/Interaction terms+in+nonlinear+models.-a0279722782
19. Williams R (2009) Using heterogenous choice models to compare logit and probit coefficients across groups. *Sociol Methods & Res* 37:531–559.
20. Greene WH (2010) Testing hypotheses about interaction terms in non-linear models. *Econ. Letters 107*.
21. Foschi M, Lai L, Sigerson K (1994) Gender and double standards in the assessments of job candidates. *Soc Psychol Quarterly* 57:326-339.
22. Bug A (2010, Aug) Swimming against the unseen tide. *PhysicsWorld.com Forum* (2).
23. Steinpreis R, Anders RK, Ritzke KD (1999) The impact of gender on the review of the CVs of job applicants and tenure candidates: A national empirical study. *Sex Roles* 41:509–528.
24. Baldi S (1995) Prestige determinants of first academic job for new sociology Ph.D.s 1985-1992. *The Sociol Quarterly* 36, No. 4:777-789.
25. National Research Council (2009) *Gender Differences at Critical Transitions in the Careers of Science, Engineering and Mathematics Faculty*. Washington DC: National Academies Press.
26. Wolfinger NH, Mason MA, Goulden M (2008) Problems in the pipeline: Gender, marriage, and fertility in the ivory tower. *J of Higher Ed* 79:388-405.
27. Glass C, Minnotte K (2010) Recruiting and hiring women in STEM fields. *J of Diversity In Higher Education* 3: 218-229.
28. Irvine AD (1996) Jack and jill and employment equity. *Dialogue*, 35 (02):255-292.
29. Kimura D (2002) *Preferential Hiring of Women*. University of British Columbia Reports.
30. Seligman C (2001) *Summary of Recruitment Activity for All Full-Time Faculty at the University of Western Ontario by Sex and Year*. At http://www.safs.ca/april2001/recruitment.html
31. Merritt DJ, Reskin BF (1997) Sex, race, and credentials: The truth about affirmative action in law faculty hiring. *Columbia Law Rev,* 97 No. 2:199-311.