2023-07

Exploring gender bias in six key domains of academic science: an adversarial collaboration

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science in the Public Interest
1–59

DOI: 10.1177/15291006231163179
www.psychologicalscience.org/PSPI


Exploring Gender Bias in Six Key Domains of Academic Science: An Adversarial Collaboration

Stephen J. Ceci[1], Shulamit Kahn[2], and Wendy M. Williams[1]
[1]Department of Psychology, Cornell University, and [2]Department of Markets, Public Policy and Law, Questrom School of Business, Boston University

Abstract
We synthesized the vast, contradictory scholarly literature on gender bias in academic science from 2000 to 2020. In the most prestigious journals and media outlets, which influence many people's opinions about sexism, bias is frequently portrayed as an omnipresent factor limiting women's progress in the tenure-track academy. Claims and counterclaims regarding the presence or absence of sexism span a range of evaluation contexts. Our approach relied on a combination of meta-analysis and analytic dissection. We evaluated the empirical evidence for gender bias in six key contexts in the tenure-track academy: (a) tenure-track hiring, (b) grant funding, (c) teaching ratings, (d) journal acceptances, (e) salaries, and (f) recommendation letters. We also explored the gender gap in a seventh area, journal productivity, because it can moderate bias in other contexts. We focused on these specific domains, in which sexism has most often been alleged to be pervasive, because they represent important types of evaluation, and the extensive research corpus within these domains provides sufficient quantitative data for comprehensive analysis. Contrary to the omnipresent claims of sexism in these domains appearing in top journals and the media, our findings show that tenure-track women are at parity with tenure-track men in three domains (grant funding, journal acceptances, and recommendation letters) and are advantaged over men in a fourth domain (hiring). For teaching ratings and salaries, we found evidence of bias against women; although gender gaps in salary were much smaller than often claimed, they were nevertheless concerning. Even in the four domains in which we failed to find evidence of sexism disadvantaging women, we nevertheless acknowledge that broad societal structural factors may still impede women's advancement in academic science. Given the substantial resources directed toward reducing gender bias in academic science, it is imperative to develop a clear understanding of when and where such efforts are justified and of how resources can best be directed to mitigate sexism when and where it exists.

Keywords
gender bias, tenure track, grants, hiring, salary/pay gap, academic letters of reference, teaching ratings, publication/productivity, women's underrepresentation in science, women in STEM, adversarial collaborations

The literature on women in science, both scholarly and popular, portrays academic sexism today as an omnipresent, pervasive force in the daily lives of tenure-track women in science, technology, engineering, and mathematics (STEM) fields. Throughout this article, we highlight quotes from prestigious journals and other outlets alleging such bias and the contexts in which it is said to occur. In response to these claims, we evaluated the scholarly evidence over a 20-year time frame—2000 to 2020—to reveal the areas of the academy in which gender bias has been addressed, as well as the areas in which it persists. First, however, we briefly describe the authors' background positions in the preface to familiarize readers with the adversarial nature of our collaboration.

Corresponding Author:
Stephen J. Ceci, Department of Psychology, Cornell University
Email: sjc9@cornell.edu

Preface

The two female authors of this article share personal histories rife with egregious examples of gender bias in academic science and beyond. Born in 1950 and 1960, respectively, they endured substantial sexism and were victims of cruelty during the earliest decades of their careers. Despite these experiences, today they share the belief—rooted in empirical data—that although the situation in academia was often deeply unfair to women in the past, it has dramatically improved over recent decades. The key question today is, in which domains of academic life has explicit sexism been addressed? And in which domains is it important to acknowledge continuing bias that demands attention and rectification lest we maintain academic systems that deter the full participation of women?

Just as there are negative consequences of not acknowledging bias, there are also costs of believing that sexism in academic science is pervasive when it is not—key among them that women will be discouraged from choosing academic careers in science, and resources will be wasted in combatting nonexistent bias claims. For this and many additional reasons, the two female authors of this article request that readers approach the topic with an open mind.

This article represents more than 4.5 years of effort by its three authors. By the time readers finish it, some may assume that the authors were in agreement about the nature and prevalence of gender bias from the start. However, this is definitely not the case. Rather, we are collegial adversaries who, during the 4.5 years that we worked on this article, continually challenged each other, modified or deleted text that we disagreed with, and often pushed the article in different directions. Although the three of us have exchanged hundreds of emails and participated in many Zoom sessions, Kahn has never met Ceci and Williams in person.

Kahn has a long history of revealing gender inequities in her field of economics, and her work runs counter to Ceci and Williams's claims of gender fairness. Kahn was an early member of the American Economics Association's Committee on the Status of Women in the Economics Profession (CSWEP). Articles of hers in the *American Economics Review* (Kahn, 1993) and in the *Journal of Economic Perspectives* (Kahn, 1995) were the first publications on the status of women in the economics profession. She was the first to identify gender inequities as a concern in economics, something she has revisited every decade since then in her publications. In 2019, she co-organized a conference on women in economics, and her most recent analysis in 2021 found gender inequities persisting in tenure and promotion in economics (Ginther & Kahn, 2021). In short, gender bias in academia has been a long-standing passion of Kahn's. Her findings diverge from Ceci and Williams's, who have published a number of studies that have not found gender bias in the academy, such as their analyses of grants and tenure-track hiring in *Proceedings of the National Academy of Sciences* (*PNAS*; Ceci & Williams, 2011; Williams & Ceci, 2015).

Although our divergent views are real, they may not be evident to readers who see only what survived our disagreements and rewrites; the final product does not reveal the continual back and forth among the three of us. Fortunately, our viewpoint diversity did not prevent us from completing this project on amicable terms. Throughout the years spent working on it, we tempered each other's statements and abandoned irreconcilable points, so that what survived is a consensus document that does not reveal the many instances in which one of us modified or cut text that another wrote because they felt it was inconsistent with the full corpus of empirical evidence.

In this article, we analyze six specific contexts—key points of evaluation for academics—in which frequent claims have arisen of bias against women scientists who are equally as accomplished as men but who nevertheless find themselves downgraded. If support is found for such bias, then it adds urgency to mitigation efforts. If support is not found, it suggests that resources might be redeployed in other ways to achieve greater benefits for women in STEM fields. There are, after all, additional aspects of academic lives beyond these six evaluation points. In particular, we do not address claims of broad societal systemic factors in this article, inasmuch as doing so would require its own monograph-length treatment.

For instance, we consider the evidence for bias in tenure-track hiring, but there is also a broader context that considers whether women reach the point of even applying for such jobs. Even if search committees treated men and women with identical curricula vitae (CVs) equivalently when they applied for tenure-track positions, women still might be impeded from applying for tenure-track positions in the first place for a host of systemic reasons, such as difficult work schedules and inflexible timing imposed by the tenure-track system during the decade when most people are building families, dominant male values in publishing style, penalizing of women for violating cultural norms when they are agentic and engage in forceful negotiating or even when they provide negative feedback in grading of students (Buser et al., 2022), and unequal childcare expectations. Reasonable people differ in their views about such broad societal construals and whether they should be called bias, and such differences exist among the authors of the present article. In contrast to these

broad societal construals, everyone agrees that if job-search committees favor male applicants over comparable female applicants, or if grant reviewers favor proposals that have male names as principal investigators (PIs), or if journal reviewers favor manuscripts that contain male names, or if the same lecture receives higher ratings when it is presented by a man, these behaviors constitute prima facie evidence of gender bias.

Introduction

A staggering number of articles has been published about women in science, reflecting research across a wide range of disciplines—psychology, sociology, economics, philosophy, biology, chemistry, physics, education, anthropology, engineering, medicine, and mathematics (more than 15,000 articles in the past decade alone).[1] It would be impossible to cover this literature in anything less than a lengthy monograph. However, our aim is not to cover it, but rather to *un*cover it—by dissecting findings across a very wide scientific landscape to achieve a synthesis.

Previous work has shown that what happens early in development can influence later gender representation in scientific careers. Such factors as early socialization differences, stereotypes, and teachers' and parents' attitudes affect students' later choices of high school advanced placement courses, which in turn are often gateways to college STEM majors and later occupations. We will not reprise here the evidence for the importance of these early social factors, because many other researchers, including ourselves, have already done so (e.g., Ambady et al., 2001; Bian et al., 2017, 2018; Carlana, 2019; Carlana & Corno, 2021; Ceci et al., 2014; Cvencek et al., 2011; Dunham et al., 2015; Lavy & Sand, 2015; Terrier, 2020).

Instead of focusing on these early systemic factors, we will focus on women who have overcome early socialization barriers to earn a STEM PhD and who are interested in tenure-track academic careers in science. We will not synthesize the research literature on non-tenure-track positions in academia (e.g., postdocs, tech/lab workers, teaching-only faculty—including adjuncts), government jobs, or jobs in industry. Nor will we discuss studies of aspects of undergraduate and graduate education. Our focus is on women eligible to compete for tenure-track positions, grants, journal acceptances, salaries, et cetera.

In short, the analysis we present in these pages is limited to claims regarding biases that women have faced in the tenure-track academy since 2000. We will not delve into pre-2000 biases and barriers that may have impeded women from vying for and succeeding in tenure-track positions. Additionally, we will not address intersectionalities of gender with race and social class because very limited data exist. Where research does exist, we will describe it, including some of our own research on race and on National Institutes of Health (NIH) grants (Ginther et al., 2016; Ginther & Kahn, 2013).

Even when limiting ourselves to a consideration of STEM tenure-track jobs, there are nevertheless many contexts in which women may face barriers to success in entering and succeeding in these jobs. In this article, we comprehensively examine evidence in six key evaluation contexts: (a) Are similarly accomplished women and men treated differently by academic hiring committees? (b) Are grant reviewers biased against female PIs? (c) Are journal reviewers biased against female authors? (d) Are recommendation-letter writers biased against female applicants for tenure-track positions? (e) Are faculty salaries biased against women? And, (f) are student teaching evaluations biased against female instructors? Claims of gender bias are omnipresent in all six of these domains (see quotes below). Our goal is to synthesize the best empirical evidence related to each of these claims to determine whether it supports the bias claims. We synthesize findings across these six evaluation contexts in an effort to reconcile contradictory results. (We also review the literature in a seventh context, gender differences in publication rates, because publishing productivity can moderate evaluation in most of these six contexts.)

These six contexts do not exhaust all possible sources of bias, even within tenure-track academia. For example, they do not cover factors such as speed of promotion, "chilly climate," speaking invitations, professional society awards, election to prestigious societies, citation imbalances, tenure rates, and persistence in the academy (e.g., Card et al., 2022; Cassad et al., 2021; Farr et al., 2017; Holliday et al., 2014; Holmes et al., 2011; Kaminski & Geisler, 2012; Mehto, 2021; National Academy of Sciences, 2007; Teich et al., 2022; Webber & Canché, 2018). Nor do these six contexts address broad systemic factors that result from underlying conditions or societal expectations that may indirectly impede women's progress, independently of whether they are treated equitably once they enter these six contexts, such as when an institution requires long working days of young mothers or when tenure schedules lack flexibility that young families need.

However, we chose these six specific contexts because they represent the most active areas of empirical research, because they are especially important early in a tenure-track career, and because they are specific enough so that gender differences in outcomes can be measured. Unlike the situation with broader

factors, virtually everyone agrees when they see evidence of these six forms of bias. If the identical CV is rated higher when a man's name as opposed to a woman's name is on it, everyone recognizes this as bias; there is no convincing counterargument. But this is not true for broad societal factors that some people regard as defensible institutional practices, such as high expectations of productivity coinciding with childbearing years for women. In such cases, the causal links demonstrating outright bias are often less compelling than is the case when a study shows that the identical grant proposal is rated higher when a man's name (vs. a woman's) is listed as PI.

Because we are interested in the current situation, we consider studies that measured the existence of bias in each of these six contexts starting in 2000. Women's roles in all aspects of society changed dramatically over the last half of the 20th century. Our research on women in academia is partially borne from earlier evidence for considerable former bias along many dimensions. Although we are interested in understanding what the current situation is, our focus is not intended to minimize or deny the existence of gender bias in the past. Thus, when we discuss each of the six contexts, where there are highly visible studies published before 2000, we note them and point out major changes over time.

Claims of gender bias in the six domains

Examples of bias claims. Claims of specific forms of gender bias in the six domains are ubiquitous, found in publications of prestigious scientific outlets such as the National Academy of Sciences and in top journals such as *Nature*, *The Lancet*, *PLOS Biology*, *PNAS*, and *Science* (e.g., Casselman, 2021; Johnson et al., 2016; Lawrence, 2006; National Academy of Sciences, 2007; Shen, 2013; Witteman et al., 2019; Witze, 2020), as well as in popular media (e.g., *Harvard Business Review*, *Wired*, *The New York Times*, *Slate*, *Huffington Post*). As a backdrop to the comprehensive empirical analysis that follows, we begin with a few narrative examples illustrating claims commonly made about gender bias in these six specific contexts. These quotes demonstrate that these claims appear in premier scientific journals and in policy statements by major scientific societies. Most of these quotes combine one or more of the six contexts that we address with references to other contexts that we do not address, illustrating the tendency to attribute bias to wide ranges of behavior. As a backdrop, consider the following quotations:

> Researchers in recent years have found that women are less likely than men to be hired and promoted, and face greater barriers to getting their work published. (Casselman, 2021, *The New York Times*, para. 9)

> Women in academia contribute more labour for less credit on publications, receive less compelling letters of recommendation, receive systematically lower teaching evaluations despite no differences in teaching effectiveness[, and] receive less start-up funding as biomedical scientists. . . . [Publications] led by women take longer to publish and are cited less often [and] are accepted more frequently when reviewers are unaware of authors' identities. . . . When fictitious or real people are presented as women in randomised experiments, they receive lower ratings of competence from scientists [and] worse teaching evaluations from students. (Witteman et al., 2019, *The Lancet*, p. 531)

> In sum, there is considerable evidence that women face persistent barriers in academia and science. (Witteman et al., 2018, bioRxiv, p. 3)[2]

> Implicit bias is pervasive. Men are preferred to women even if they have the same accomplishments. Psychologists have shown this by testing scientists' responses to fictitious CVs that are identical other than coming from "John" or "Jennifer." (Witze, 2020, *Nature*, p. 25)

> A vast literature . . . shows time after time, women in science are deemed to be inferior to men and are evaluated as less capable when performing similar or even identical work. This systemic devaluation of women results in an array of real consequences: shorter, less praise-worthy letters of recommendation; fewer research grants, awards, and invitations to speak at conferences; and lower citation rates for their research. Such wide-ranging devaluation of women's work makes it harder for them to progress in the field. . . . These are just a few of the hundreds of peer-reviewed studies that clearly show, on average, the bar is set higher for women in science than for their male counterparts. (Coil, 2017, *Wired*, paras. 3 and 5)

> Considerable research has shown . . . evaluation criteria contain arbitrary and subjective components that disadvantage women. (National Academy of Sciences, 2007, p. 3)

> Women have fewer publications and collaborators and less funding, and they are penalized in hiring decisions when compared with equally qualified

> men. (Fortunato et al., 2018, *Science*, p. 3, citations omitted)

Why discuss absence of bias? Some readers might ask why it is important to discuss situations in which there is no evidence of gender bias, as well as those in which there is evidence. We believe that there are three major reasons. First, identifying areas in which bias no longer exists allows the research community to focus its efforts on career aspects, junctures, and systems that continue to disadvantage women (e.g., lengthy periods of postdoctoral study). Second, as in the story of the boy who cried wolf, if people claim unfair bias every time there is outcome asymmetry—and if the well-intentioned efforts of many academic leaders and administrators, teachers, policymakers, and funders over the past 20 years to equalize treatment of scientists irrespective of gender appear not to have improved the situation—many people may give up in discouragement (which would be particularly regrettable if their efforts actually worked). Third, if women erroneously believe that every aspect of academia is biased against them, some STEM PhDs may be reluctant to enter any area of academia, even those that are biased in favor of women, and may instead seek jobs in industry or government. Some women may simply not consider a research career in STEM because they fear working in a sexist environment, even when data might show that this environment is not biased against them.

Again, we emphasize that, in what follows, we focus on the evaluation of gender bias in academic tenure-track jobs, which are often seen as the pinnacle of employment, well-paid and secure. Perhaps more importantly, faculty in these positions educate the next generation of scientists. Although there are claims of bias outside the tenure-track academy (e.g., among lecturers or in industry, such as by Ding et al., 2021), any assessment of such claims is outside of our objective.

As we review the empirical evidence related to each of the six tenure-track evaluation contexts (as well as the seventh domain of productivity), we will highlight analyses addressing causal factors. These include randomized experiments, naturally occurring events that provide opportunities to test impacts under investigation, and multivariate analyses controlling for confounding factors. Our goal is to determine whether these analyses lead to the same conclusion. In all, we discuss hundreds of studies—some of which are meta-analyses of hundreds of other studies—exploring how fairly women are evaluated. Thus, we cover a very wide swath of research.

We strove to include studies that were sound methodologically, even if they contradicted each other or the personal views of some team members.[3] When appropriate, we conducted and report on our own meta-analyses in domains in need of them. When we did not review all studies in a given domain, we provide a principled reason. (See Kahn et al., 2022a, 2022b, for specification of the meta-analyses' search terms, inclusion criteria, Preferred Reporting Items for Systematic Reviews and Meta-Analyses [PRISMA] diagrams, funnel plots, forest charts, and related technical information that guided our searches and informed our meta-analyses.) Finally, we based our analyses and conclusions on the actual data published in these articles, which sometimes diverged from interpretations that the authors or others made about these data.

Context matters. Bias occurs within a context. Not all scientific fields are equal in their representation of women (Ceci et al., 2014; Cheryan et al., 2017), and therefore, they may not be equivalent in their evaluation of women. Similarly, as we show below, granting agencies in different countries have very different gender gaps in success rates. When the data show that differences exist across scientific fields, professorial ranks, and nations, we are careful to point out these differences.

Historical background on women's participation in science

Historically in the United States, men were more likely to earn baccalaureates than women. In 1900, only 19% of college degrees went to women. This number rose steadily throughout the century, except during the aftermath of World War II (influenced by the 1944 GI Bill, 41% of bachelor of arts [BA] recipients were female in 1939, but only 27% were female in 1950 because of the influx of World War II veterans).[4] By 1982, women received slightly more than 50% of BAs, and this rose to 57% by the turn of the century, after which the percentage of BAs awarded to women stabilized. Women's educational ascendancy does not stop with the baccalaureate: They also earn more master's degrees (59%) and more PhDs (53%) than men.

Notwithstanding the progress made by women, in math-intensive fields, they have a much lower level of representation than men. Figure 1 empirically demonstrates this, showing women's representation among baccalaureate, PhD, and tenure-track appointments. In this graph, we divided two national data sets into the most mathematically intensive fields—geosciences, engineering, economics, mathematics/computer science, and physical science (GEMP)—and less math-intensive fields—life sciences, psychology, and social sciences (LPS). Because of these differences in women's representation, it is common to distinguish between GEMP and LPS fields, and we do so here when possible.

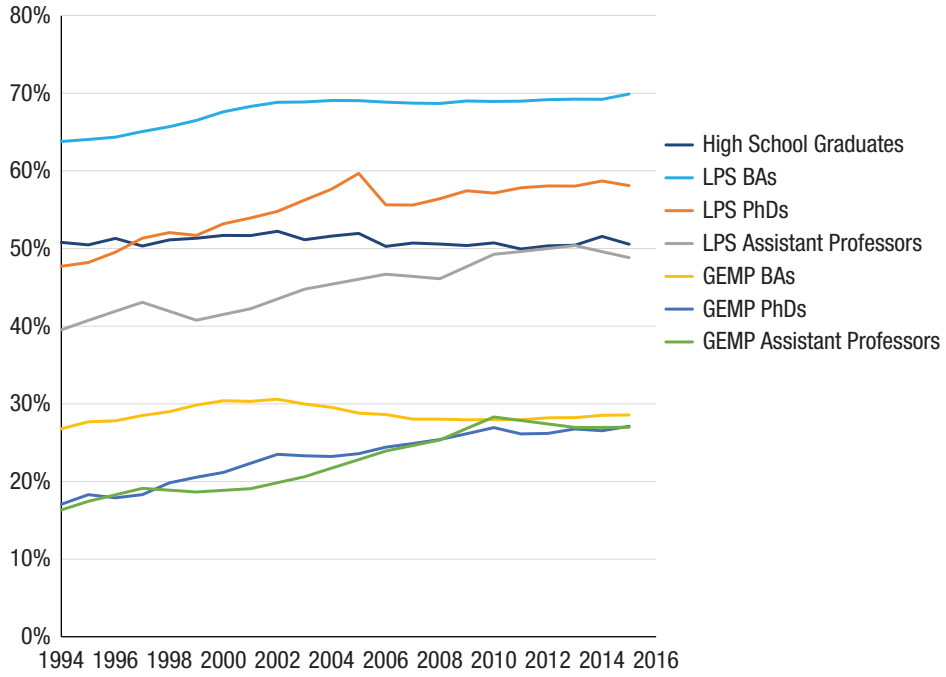

Fig. 1. Percentage of female high school graduates, bachelor of arts (BAs), PhDs, and tenure-track assistant professors by major field from 1994 to 2016. GEMP = geoscience, engineering, economics, mathematics, computer science, and physical science; LPS = life sciences, psychology/behavioral sciences, and social sciences. Source for baccalaureate degrees: WebCaspar (https://ncsesdata.nsf.gov/webcaspar/); source for assistant professor data: National Science Foundation's Survey of Doctorate Recipients (https://www.nsf.gov/statistics/srvydoctoratework/).

The dearth of women in GEMP fields has been attributed to many factors, such as mathematical ability, mathematical education, and people-vs.-things orientation (Su & Rounds, 2015; Su et al., 2009). Litigating these alleged reasons is not our focus here. For interested readers, Ceci et al. (2014) provide a discussion of these factors.

Below, we consider claims of gender bias in six evaluative contexts, as well as the potentially important and cross-cutting influence of productivity—for example, the role journal publications might play in hiring, receipt of grants, recommendation letters, and salary decisions.

A Search for Gender Bias in Six Evaluation Contexts

Evaluation Context 1: tenure-track hiring

High-profile claims are often made that STEM hiring is biased against women. This includes the quotes by Witze (2020) in *Nature* and Casselman (2021) in *The New York Times* cited earlier. Many similar claims of gender bias in tenure-track hiring have been made. Consider the following quotations:

> Research has pointed to bias in peer review and hiring. . . . A female 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. (Hill et al., 2010, p. 24)

> Even after earning STEM degrees, women are less likely to be hired into STEM jobs compared with equally qualified men. (Cech & Blair-Loy, 2019, p. 4182)

A few studies examined the presence of racial or gender bias in the evaluation of résumés and CVs (e.g., Moss-Racusin et al., 2012) or in recruitment processes (Milkman et al., 2015; Posselt, 2016), suggesting that the lack of faculty diversity can be attributed to bias in institutional gatekeeping processes, such as hiring (O'Meara et al., 2020). However, none of these were tenure-track-hiring studies, and later we show that there are important reasons for not assuming that such studies inform gender bias in the tenure-track academy.

Because of the heterogeneity of dependent variables, forms of evidence, nations, epochs, ranks, type of institutions, and disciplines, meta-analysis in the tenure-track-hiring domain is inappropriate. Fortunately, the findings and effect sizes from the myriad analyses strongly point in the same direction, as we will show. In synthesizing these findings, we will examine three types of evidence that have been invoked in support of the gender-bias argument: (a) cross-sectional, (b) cohort, and (c) experimental. Below, we provide systematic coverage of the empirical literature addressing two of these three types of evidence.

Cross-sectional comparisons. Many claims of bias in tenure-track hiring are based on contemporaneous cross-sectional percentages of PhDs, tenure-track untenured professors, tenured associates, and full professors. An example illustrates this:

> Between 1969 and 2009, the percentage of doctorates awarded to women in the life sciences increased from 15% to 52%. Despite the vast gains at the doctoral level, women still lag behind in faculty appointments. Currently, only 36% of assistant professors and 18% of full professors in biology-related fields are women. The attrition of women from academic careers . . . undermines the meritocratic ideals of science and represents a significant underuse of the skills that are present in the pool of doctoral trainees. (Sheltzer & Smith, 2014, p. 10107)

However, given the growth in the percentage of female PhDs in life sciences over these decades, we should not expect the percentage of female full professors in 2009 (most of whom received their PhDs from 1980 to 1995) to be 52%. Expectations of the percentage of female full professors must be based on the relevant cohorts rather than on contemporaneous PhD numbers. In another example, the European Commission, Directorate-General for Research and Innovation (2019) concluded, "Women face greater difficulties than men in advancing to the highest academic positions in all the countries examined" (p. 125). Once again, this conclusion was linked to the number of newly minted PhDs rather than to the relevant cohort. In a third example, Golbeck (2016) used cross-sectional data to argue that women were underrepresented as tenure-track professors in 2014, pointing out that they comprised 46% of new PhDs but only 33% of tenure-track assistant professors—a 13 percentage points (ppt) gap. However, using the appropriate 2014 cohort, we found that female assistant professors in 2014 comprised 40% of PhDs but only 37% of tenure-track assistant professors, a 3 ppt gap. This gender gap was even smaller or reversed when we adjusted for cohort because 2008 is the oldest year in which candidates could receive PhDs and still be on the tenure track in 2014, and women comprised only 31% of PhDs that year. This erases Golbeck's estimate of a 13% gender gap.

Thus, these gender differences need not indicate a bias in the hiring of women with newly minted PhDs, for two reasons. The first reason is that in fields where women constitute steadily increasing proportions of PhDs each year, cross-sectional contemporaneous comparisons will always overestimate gender differences in the probability of proceeding to tenure-track jobs at a given point. This is not a new insight (see, e.g., Abramo et al., 2021; Hargens & Long, 2002; Stewart et al., 2009), yet many scholars still fail to take this issue into account. The next section presents evidence on whether in the United States, PhD women graduates progressed into academia at a similar pace as men. The second reason that gender differences in the likelihood of PhDs entering tenure-track jobs need not imply bias is that women PhDs are less likely to apply for tenure-track positions. We address evidence for this in the following section. We then describe studies that do address whether there is gender bias in the likelihood that PhDs who apply for tenure-track jobs are hired.

Evidence on transitioning from PhD to tenure-track positions based on actual or synthetic U. S. cohort comparisons. To compare women's and men's transition from PhD to tenure track, one can use actual cohorts (longitudinal data); "synthetic cohorts," matching professors to the actual PhD cohorts feeding them; or counterfactual hypothetical candidates from the relevant cohort. For the United States, we created synthetic cohorts using the National Science Foundation's (NSF's) Survey of Doctorate Recipients (SDR), a biennial survey of 120,000 doctoral degree holders conducted by the National Center for Science and Engineering Statistics (NCSES) within NSF. The SDR is the best source of national longitudinal data on PhD recipients in science and health fields. In these analyses, we juxtaposed the annual supply of new PhDs in GEMP and LPS fields, respectively, with the assistant professors in the years during which they would be expected to be at that rank if they entered tenure-track academia.

For separating GEMP and LPS fields by year, we first calculated the smallest range of PhD years of at least 50% of tenure-track assistant professors that year. (For instance, 50% of 1993 assistant professors received PhDs between 1985 and 1990.) With this information, we calculated the percentage of women who received PhDs each year and the percentage of female assistant professors we would expect to see 5 to 8 years later if men and women were equally likely to progress from PhD to assistant professor (see Fig. 2).

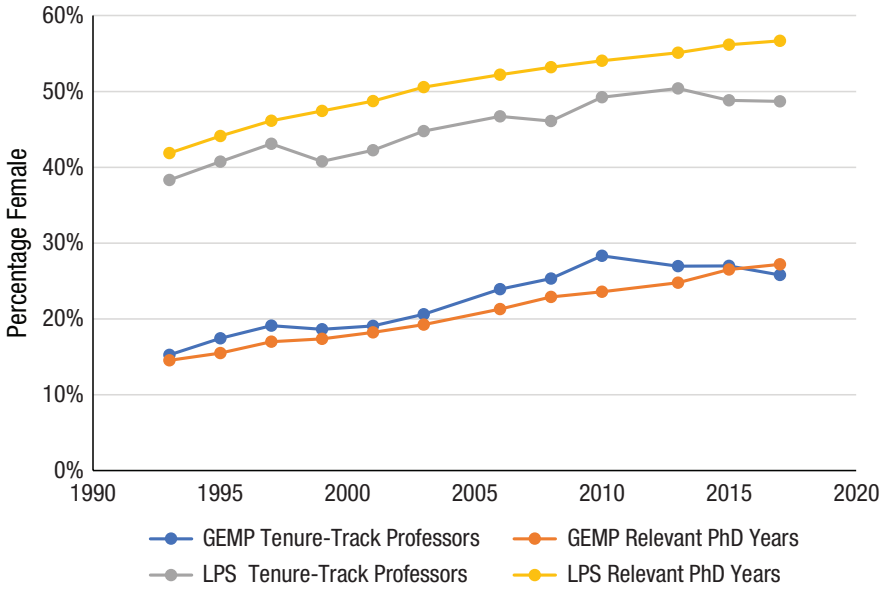

Fig. 2. Percentage of female tenure-track assistant professors from 1994 to 2017 and women in the relevant PhD cohorts by major field. Calculations were made from data obtained from the National Science Foundation's Survey of Doctorate Recipients (https://www.nsf.gov/statistics/srvydoctoratework/). Relevant PhD years are the period when the middle 50% (25th–75th percentile) of the assistant professors in the corresponding year graduated. GEMP = geoscience, engineering, economics, mathematics, computer science, and physical science; LPS = life sciences, psychology/behavioral sciences, and social sciences.

Figure 2 shows that for fields in which women are most underrepresented—the math-intensive GEMP fields—the percentage of female assistant professors is either approximately the same as the percentage of female PhDs in the relevant feeder years for that cohort or is greater (as in 2010). The prima facie evidence, therefore, suggests that female applicants for GEMP tenure-track positions have been slightly more likely to be hired than men since 1993. This female advantage, while small, runs counter to the common claim that women are far less likely to be hired as a result of bias.

However, the pattern is different in the LPS fields, despite women's overall representation being much greater than in GEMP fields: The percentage of female tenure-track assistant professors in LPS is always lower than the percentage of female PhDs in the relevant feeder years. The gap was smallest in the mid-90s—less than 4 ppt—growing to 8 ppt in 2017. The percentage of female tenure-track assistant professors in LPS fields has hardly changed since 2010, hovering between 49% and 50%, despite increases each year in the feeder pool of female PhDs.

Using actual or synthetic cohort analyses, other researchers have also found less pronounced gender gaps or even overrepresentations of women in GEMP fields. Comparing the percentage of PhDs conferred to women between 1996 and 2005 with faculty in 2007, Kessel and Nelson (2011) reported that female PhDs had similar or higher probabilities than men of entering assistant professorships in 100 top "highly quantitative" departments but not in other STEM fields. Ceci et al. (2014) compared the percentage of female PhDs with the percentage of female assistant professors 5 to 6 years later in GEMP fields and found similar results. And in philosophy—the humanities field most like GEMP in gender composition and quantitative emphasis—among 2008 to 2019 PhDs, women had a 10% to 17% greater likelihood than men of entering permanent academic placements (Allen-Hermanson, 2017; Kallens et al., 2022).

Thus, these cohort analyses offer little support for the claim of widespread gender discrimination in tenure-track hiring in GEMP, even before 2000. Economics is the exception; it is the only GEMP field in which analyses of synthetic cohorts indicate otherwise. CSWEP creates synthetic cohorts annually (e.g., Chevalier, 2019), showing that the percentage of women among tenure-track assistant professors (within 7 years after obtaining their PhDs) was similar to the percentage of women among PhDs only through 2004; for the next eight PhD cohorts, however, the percentage of female assistant professors stagnated, despite growth of newly minted female PhDs.

Two studies used the longitudinal capability of NSF's SDR. Wolfinger et al. (2008) found that among PhDs

from 1981 to 1995, women were 7% less likely than men to transition to the tenure track. However, using a longer range of the SDR and differentiating among fields, Ginther and Kahn (2009) found that among people who earned PhDs between 1973 and 2001, women and men in physical sciences and engineering were equally likely to transition to the tenure track, whereas in life sciences, women were 7.7 ppt less likely. For social sciences, Ginther and Kahn (2014) found that women were 3.7% less likely. Although these analyses all controlled for PhD department ranking, they did not control for publication productivity or for the aspiration for tenure-track positions, which differs between women and men, as we will show later. (Interestingly, both Wolfinger et al., 2008, and Ginther and Kahn, 2009, found that the largest female demographic of job hunters—unmarried single women—were 15% more likely than men to transition to the tenure track.)

Finally, an older audit of the hiring of assistant professors across all nine University of California campuses between 1990 and 1999 compared the percentage of women among those hired as assistant professors with the percentage of women among recent PhDs awarded in the University of California system. Women were overrepresented among new hires in four of the six GEMP fields: In physics, the percentage female among hired assistant professors was higher than among recent PhDs (women were hired 14% of the time vs. 10% female PhDs); in engineering, women were hired 13% of the time versus 8% female PhDs; in geoscience, women were hired 28% of the time (vs. 19% women PhDs); and in computer science, women were hired 14% of the time (vs. 10% female PhDs). However, in two GEMP fields, there was a pronounced male advantage: In mathematics, women were hired as assistant professors far less often than the numbers of recent female PhDs (women hired 7% of the time vs. 20% female PhDs), and in chemistry, women were hired 7% of the time (vs. 26% PhDs; California State Auditor, Bureau of State Audits, 2001, Table 2).

Female PhDs are less likely to apply for tenure-track positions. However, even if men and women in the same cohort had proceeded at different rates to the tenure track, this would not prove bias in hiring because not every PhD holder aspires to or applies for tenure-track jobs. Surveys in both the United States and Europe consistently show that women are significantly less likely than men to pursue a tenure-track job as they advance through training (e.g., Bataille et al., 2017; Blinkenstaff, 2005; Salinas & Bagni, 2017; Schubert & Engelage, 2011; Trevino et al., 2017).

Studies suggest that the lower rate of applying for tenure-track positions is the result of systemic social/structural and biological factors, such as women's disproportionate family, childbearing, and child-rearing responsibilities combined with the rigid time frame of the tenure process. For instance, in the United States, women are twice as likely as men to abandon their careers after childbirth (Cech & Blair-Loy, 2019; Skibba, 2019). Goulden et al. (2009) found leakage from research careers among postdocs to be especially apparent among women with children or planning to have children (28% leakage for women vs. 16% for men), whereas women without children or plans to have children were as likely as men to pursue the tenure track. Goulden et al. (2009) found that the gender gap in applying for faculty positions was completely accounted for by family formation decisions, and Martinez et al. (2007) found in a survey of 1,300 NIH postdocs that 21% of women compared with 7% of men said that plans to have children or additional children were extremely important in planning research careers. Finally, Ecklund and Lincoln's (2011) survey of 3,455 biologists, astronomers, and physicists in top-20 departments found that 4 times as many female graduate students and 50% more female postdocs worried that a science career would keep them from having a family. We will revisit this important point of leakage of women below.

We saw that in LPS fields, women PhDs were less likely to proceed to tenure-track jobs than men. Is the lower likelihood of applying particularly marked for women in these fields? It would make sense, given that postdocs are usually required in biology and are increasingly common in psychology, to enable a chance of getting tenure-track jobs, making them even less attractive to PhD women thinking of starting a family. We can see in the National Research Council (NRC; 2010) data shown in Table 1 that the percentage female of tenure-track applicants is only 58% of the percentage female of PhDs in biology, and that this is much smaller than the percentage for engineers, mathematicians, or physicists. Relatedly, Martinez et al. (2007) also found extremely different career plans for men and women in biomedical fields. And recent research by Cheng (2020) also found that in biological sciences, mothers with children are much less likely to enter or remain in tenure-track jobs.

Such leakage is likely the result of systemic challenges posed by inflexible tenure-track jobs, coupled with women's and men's differing experiences with and expectations of childbearing and parenting. However, it says nothing about bias in the tenure-track hiring process, which may be gender fair for those women who do apply. Thus, it is important to control for actual numbers of women applying for tenure-track jobs rather than assume that equal percentages of men and women who earn PhDs vie for tenure-track jobs. The latter is simply not true.

Table 1. Percentage of Women Among Applicants, Interviewees, and People Offered Positions in 545 Tenure-Track Searches

	Transitions for PhD-to-tenure-track positions by field at the research institutions surveyed			
	Doctoral pool	Pools for tenure-track positions		
Field	Percentage of female PhDs (1999–2003)	Mean percentage of female applicants	Mean percentage of female applicants invited to interview	Mean percentage of first offers that went to women
Biology	45	26	28	34
Chemistry	32	18	25	29
Civil engineering	18	16	30	32
Electrical engineering	12	11	19	32
Mathematics	25	20	28	32
Physics	14	12	19	20

Note: Data were obtained from National Research Council (2010, Table S-2, p. 7).

To recap, despite claims of gender bias in tenure-track hiring, our national cohort analyses show no increased likelihood that men proceed to tenure-track jobs relative to women in the very fields in which women are most underrepresented (GEMP), although there is a difference in LPS fields. A major factor—if not the only factor—responsible for women being less likely to make this transition is that they do not apply as often for these jobs. Neither of these facts, however, directly answers the question, "If an equally qualified man and women apply for a tenure-track job, will the man be more likely to get it?" Below, we describe three kinds of studies that directly address this question: (a) cohort studies limited to graduates who entered academia, (b) large-scale institutional analyses, and (c) experimental evidence. These three categories of evidence point in the same direction, which suggests gender-fair or female-preferred hiring.

Cohort analyses limited to people likely to apply. Two studies limit their analysis of hiring to graduates who applied for academic jobs by including only those women and men who did compete for academic jobs at some point and by measuring whether women were hired into more prestigious departments. In the United States, two studies limited their analyses to women entering academic jobs. In computer science, Way et al. (2016) found that more highly ranked departments hired women and men at comparable rates, holding constant publications, department prestige, geography, and postdoc experience. An earlier study by Clauset et al. (2015) used the same methodology on more fields but had no information on applicants' publications or postdocs: This study found that women tended to be hired by lower-ranked departments than men were. Combining these two findings suggests that publications and postdocs accounted for the entirety of the tenure-track hiring gender gap in prestigious departments.

Two other studies relate to German academia. Germany's academic system is different than that in the United States. There are no entry-level tenure-track jobs, only (national) competitions for tenured professorships (the only academic rank with permanent appointments). Schröder et al. (2021) analyzed the entire population of German political science departments in 2018 or 2019, excluding individuals with PhDs received before 1980. (These included current predocs, postdocs, temporary faculty, and professors.) Lutter and Schröder (2016) did the same for individuals in sociology departments in universities in 2013. They limited the analysis to only individuals who published at least once and also only to those observed at that point in time in a university. These two criteria probably eliminated a large proportion of PhDs not interested in academic research jobs.

Among political scientists, Schröder et al. (2021) found that female political scientists had a 20% greater likelihood of obtaining a tenured position than comparably accomplished males in the same cohort after controlling for personal characteristics and accomplishments (publications, grants, children, etc.). Lutter and Schröder (2016) found that women needed 23% to 44% fewer publications than men to obtain a tenured job in German sociology departments. As Schröder et al. (2021) concluded,

> Ours is the first study to use a virtually complete sample of all German academic political scientists to show that women tend to be favored over men in the hiring process for tenured professorships,

> before and after controlling for various factors, most importantly productivity (but see [Lutter & Schröder, 2016] for the discipline of sociology in Germany). This means that women get hired with fewer measurable publications than men do, indicating that there is no bias against women when judging their competency, different from what other studies found. (para. 52)

Evidence of bias based on actual university hiring data. Audits of actual university hiring, which often occur for affirmative action/inclusion purposes, can establish whether women who apply are less likely to be hired than men. Generally, this evidence consists of institutional data that are not posted online or accessible outside the university or available through online search engines. Below, we describe seven such reports that were published or temporarily available online before being taken down.[5]

This subsection is not based on a comprehensive search but rather on institutional findings that have been made available, and therefore they may entail biases associated with a university's willingness to make its administrative data available: Institutions that discover evidence of employment discrimination in either direction may be less likely to make such reports discoverable. On the other hand, the NRC (2010) study of approximately 500 departments in 89 of the most prestigious research-intensive (R1) universities in the United States had an 85% response rate, avoiding the above limitations. Also, some of these reports are dated, based on hiring during the 1970s to 2000. Thus, we offer the evidence in this subsection with caution, and readers may in fact wish to de-emphasize this subsection even though it is consistent with the findings from the cohort analyses described above and the experimental analyses that follow.

One of the largest sources of hiring audit data is an NRC (2010) study of 545 tenure-track hires and 97 tenured faculty hires from 1995 to 2003 at 89 R1 U.S. universities in five GEMP fields plus one LPS field (biology). As was true of the studies discussed above, this NRC report found, "The percentage of applications from women are consistently lower than the percentage of PhDs awarded to women" (p. 48). As shown in Table 1 (which reports data from Table S-2 of NRC, 2010), they found differences between disciplines and specifically that, "in the fields with the largest representation of women with PhDs—biology and chemistry—the percentage of PhDs awarded to women exceeds the percentage of applications from women by a large amount" (pp. 47–48).

This table also shows that in all disciplines, women who applied for tenure-track jobs were invited to interview and were offered positions at rates higher than men were, leading the NRC team to conclude that

> women fared well in the hiring process at Research I institutions, which contradicts some commonly held perceptions of research universities. If women applied for positions, they had a better chance of being interviewed and receiving offers than male job candidates had. (NRC, 2010, pp. 4–5)

Table 2. Results From a Study of Applicants for Tenure-Track Jobs in 19 Scientific Fields at a Large Research University

		Likelihood of proceeding to next stage (%)	
Group	Women (%)	Men	Women
Applicants	15.2	32.7	46.4
Offered	26.3	72.6	76.9
Hired	27.4		

Note: Adapted from Glass and Minnotte (2010, Table 1).

Similar results from actual university hiring data were reported for a later period (2000–2005) by Glass and Minnotte (2010), who studied 3,077 applicants for 63 tenure-track jobs in 19 scientific fields at a large research university. We calculated the percentages shown in Table 2 from their data. Again, women's likelihood of being hired was greater than men's.

There are two published Canadian reports of hiring that, although older, accord with the findings above. At the University of Western Ontario, across departments in 1992 to 1999, women constituted 23.2% of applicants, 30.4% of interviewees, and 36.2% of hires for tenure-track jobs (The University of Western Ontario, Office of the Provost and Vice-President [Academic], 2001). At Simon Fraser University and University of British Columbia in 2001, of 4,525 applicants, women were more likely than men to be one of the 105 hired, comprising 38.9% of applicants but 41.0% of those hired (Kimura, 2002).

Another example of university administrative hiring data is an online report of 5 years of hiring at the University of California, Davis (November 2009–October 2014). Table 3 shows increasing percentages of women as we move through the hiring process. By our calculation, 31.0% of the University of California, Davis applicants were women, 36.5% of those interviewed were women, and 40.7% of those hired were women. As was true in the NRC (2010) findings, there was considerable heterogeneity, but in all fields except biology, the percentage of women hired surpassed the percentage of female applicants. The biology results contrast with the NRC findings, however.

Table 3. Percentage of Women Who Applied, Were Interviewed, and Were Hired for University of California, Davis Academic Positions (November 2009–October 2014)

Field	Applied (%)	Interviewed (%)	Hired (%)
Agricultural/environmental science	32.4	30.2	42.5
Biological science	37.6	34.0	27.3
Engineering	15.0	20.0	25.0
Math and physical science	15.2	31.4	23.1
Social science	34.0	43.0	38.5
Medicine	25.2	34.2	46.8
Veterinary medicine	23.1	28.6	34.8
Humanities	45.0	51.0	57.1
Education	62.3	63.2	62.5
All (including law, business, education, and humanities)	31.0	36.5	40.7

Note: Data were recovered from the University of California, Davis public web page but are no longer posted.

The same overall picture appears in administrative data from other countries. An analysis by Moratti (2020) of hiring for the decade from 2007 to 2017 at Norway's largest university revealed no gender bias in hiring: Seventy-seven searches generated 1,009 applicants for new associate professorships, with women slightly more likely than men to be hired, leading Moratti to conclude that

> women applicants between 2007 and 2017 had a slightly higher likelihood of victory compared to their men competitors. . . . Much of the international literature . . . on application patterns and selection outcomes for permanent academic positions reports that women are either reluctant to apply or systematically dispreferred (or both); that is not what we found. (pp. 924 & 927)

These analyses of institutional hiring data are consistent with a national audit of computer-science hiring showing that women were more successful in obtaining offers of faculty positions. The Computer Research Association commissioned a national audit of U.S. and Canadian computer-science hiring (Stankovic & Aspray, 2003). They found that new women recipients of PhDs applied for far fewer academic jobs than men: Women with PhDs applied for six positions, whereas men applied for 25 positions. However, female PhDs were offered twice as many interviews per application (0.77), whereas men received only 0.37. Further, women received 0.55 job offers per application, whereas men received only 0.19: "Obviously women were much more selective in where they applied, and also much more successful in the application process" (Stankovic & Aspray, 2003, p. 31).

In summary, all of the seven administrative reports reveal substantial evidence that women applicants were at least as successful as and usually more successful than male applicants were—particularly in GEMP fields. (This is also the case in the eighth study we chose to delete because its authors have not made their data publicly available yet.) Several of these reports also found that women PhDs were less likely to apply to these jobs than men. This conclusion is ratified by Salinas and Bagni's (2017) review of the underrepresentation of women scientists in European studies:

> Analyses of the female candidates applying for independent positions suggest that [underrepresentation of women] is not due to discrimination against women, but rather to the fact that fewer women apply for jobs as independent investigators. . . . Once a female applicant enters the recruitment process, she has an equal chance to be offered the position compared to her male counterparts. (p. 722)

To recap, both the cohort analyses that are limited to individuals who likely applied for academic jobs and the seven university administrative hiring analyses just reviewed point to gender-fair or even pro-female hiring. Moreover, even without limiting analyses to applicants or likely applicants, women PhDs in GEMP fields are just as likely as men to proceed to tenure-track jobs. On the other hand, the cohort analysis for women PhDs in LPS fields found that women were less likely than men to proceed to tenure-track jobs. However, there is some evidence that women in biology are particularly less likely to seek tenure-track jobs.

Unfortunately, the audits of administrative hiring data presented above are limited to what is publicly available, and much of it is at least 15 years old. It is possible—even likely—that similar data exist at other universities but are not publicly posted. Despite

numerous search strategies, we were unable to unearth these data. We hope that studies similar to these are pursued to add to the results to date.

Below, we discuss the final source of hiring information—experimental studies. Overall results accord with the same gender-fair (or pro-female) interpretation seen in the administrative and cohort findings just reviewed.

Experiments in hypothetical hiring. The above large-scale analyses of nationally representative cohort data, as well as institutional data, document gender-neutral or pro-female hiring but not its cause. To explore whether women are favored in tenure-track hiring because they are stronger candidates (e.g., women's greater attrition during graduate school could theoretically result in those who survive being stronger and having more publications) or for other reasons (e.g., desire for diversity), some researchers have conducted experiments or quasi-experiments in which identically qualified men and women vie for jobs. These studies do not mirror hiring by search committees because (a) when faculty in experiments evaluate hypothetical applicants, they may engage in virtue signaling because nothing is at stake, and (b) actual hiring is influenced by discussions among committee members or entire departmental faculties, and such conversations are missing from hiring experiments. On the other hand, experiments provide a means of testing many hypotheses, such as whether variations in a given factor (e.g., one extra publication or the presence of children) affect women's and men's hirability similarly.

We conducted an exhaustive search on Web of Science and Google Scholar of studies of CV-matched tenure-track hiring experiments. Only four such studies met our inclusionary criteria of being tenure-track experiments in the 2000 to 2020 period (and a fifth was from 1999). We supplemented these searches with a bottom-up search of every individual cite to each of these five studies (> 1,300 cited studies). No other experimental tenure-track study emerged during the past 20 years, save one that was published a few months after the closing of the inclusion period (Henningsen et al., 2021).

We begin with a 1999 experimental study of tenure-track hiring, even though it fell 1 year outside our inclusionary date. We decided to include it because it had more Google Scholar cites (913 through 2021) than all of the other four studies combined, and it is the sole study to show results in the opposite direction from our conclusion. For these reasons, including it seemed important. Steinpreis et al. (1999) sent 238 academic psychologists one of two hypothetical CVs of a fictional scientist applying for a new tenure-track post, with either a female or a male name. Male candidates were more likely to be recommended for a tenure-track job by both male and female psychologists. (Steinpreis et al. also studied promotion to tenure for two identically qualified advanced candidates but found no gender difference.)

Williams and Ceci (2015) studied a stratified national sample of 872 faculty from two GEMP fields (engineering and economics) and two LPS fields (biology and psychology) to determine preferences for identically qualified men and women possessing outstanding credentials. Figure 3 shows that in the authors' main experiment ($N = 363$), faculty expressed a significant preference for hiring women. This pro-female preference was similar across fields, types of institution, and gender and rank of faculty. The only group in which the preference did not appear was male economists, who showed no gender preference. The difference between the findings of Steinpreis et al. (1999) and Williams and Ceci (2015) may be due to differences in the strength of the hypothetical candidates (see Ceci & Williams, 2015), although both studies described them as being very strong, or perhaps to changing faculty attitudes toward gender diversity over the intervening 16 years. Steinpreis et al.'s data were collected in the mid-1990s, and differences may be due to changes over time.

In a natural experiment, French economists used national exam data for 11 fields, focusing on PhD holders who form the core of French academic hiring (Breda & Hillion, 2016). They compared blinded and non-blinded exam scores for the same men and women and discovered that women received higher scores when their gender was known than when it was not when a field was male dominant (math, physics, philosophy), indicating a positive bias, and that this difference strongly increased with a field's male dominance. Specifically, women's rank in male-dominated fields increased by up to 40% of a standard deviation. In contrast, male candidates in fields dominated by women (literature, foreign languages) were given a small boost over expectations based on blind ratings, but this difference was small and rarely significant.[6]

Carey et al. (2020) conducted an experiment with 869 faculty at two U.S. state universities. They presented participants with two profiles of hypothetical candidates to be hired as faculty members. Each profile included several attributes relevant to the hiring decision, one of which was randomly selected per faculty respondent and the order of which was randomized across participants. This resulted in a conjoint analysis that allowed for the calculation of marginal component effects to reveal the combination of attributes that faculty rely on to make their faculty-hiring decisions. The authors found that all else being equal, faculty were between 5% and 10% more likely to favor a female candidate or a gender nonbinary candidate, respectively, over an identically accomplished male.

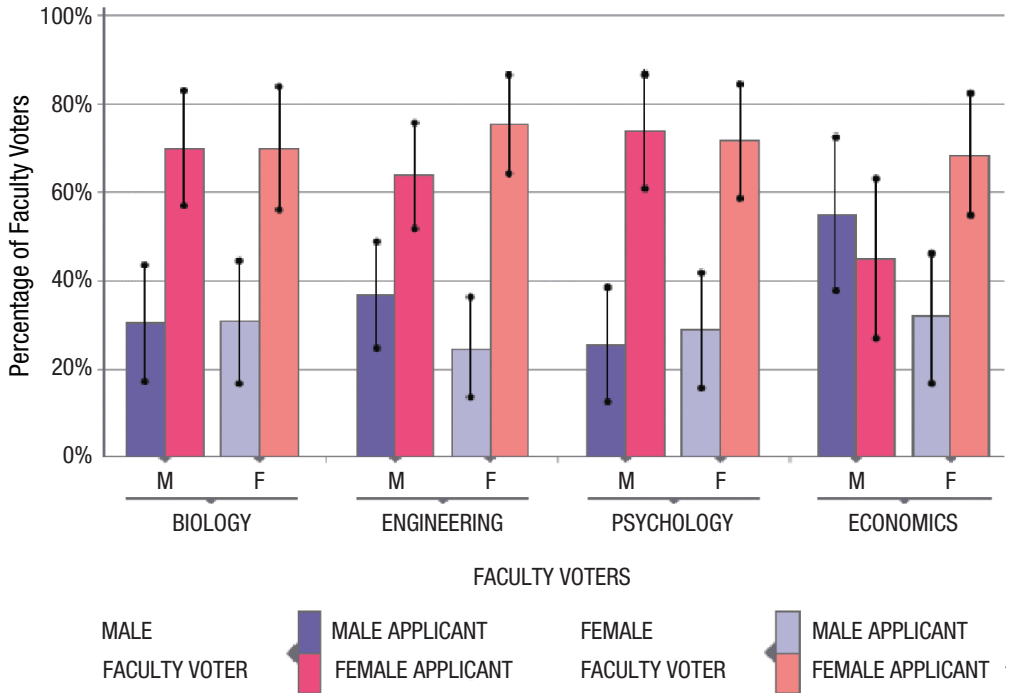

Fig. 3. Percentage of faculty members who ranked male (M) and female (F) applicants as the top candidate, separately for male and female faculty voters in each of four disciplines. Error bars indicate ±1 *SE*. Adapted from Williams and Ceci (2015, Experiment 1).

Finally, a large-scale experiment by M. Carlsson et al. (2021) using faculty at 17 large institutions in Iceland, Sweden, and Norway also revealed a pro-female hiring bias. Their experimental design most closely resembles that of Williams and Ceci (2015), who found a significant female advantage, although Carlsson et al. used a between-subjects manipulation rather than the within-subjects manipulation that Williams and Ceci used in most of their experimental conditions. On the basis of actual CVs, Carlsson et al. prepared experimental CVs that varied the hypothetical applicant's gender, marital status, family status (children vs. no children), and research productivity (four vs. six international publications). Out of a target population of 2,000 faculty in STEM fields plus law, 775 faculty agreed to participate (39%). The authors found a significant pro-female advantage, with faculty rating female applicants' competence and hirability significantly higher than identically accomplished male applicants'. This can be seen in the nonoverlapping competence and hirability ratings in Figure 4.

These findings contradicted several of M. Carlsson et al.'s (2021) preregistered hypotheses, leading them to conclude as follows:

> We expected our survey experiment to reveal a male advantage. . . . We also expected female candidates to have a lower return to children and strong[er] CVs than males. . . . Contrary to our main hypothesis, however, [our analyses showed that] female candidates are perceived as both more competent and hireable compared to equally qualified male candidates. Furthermore, we find no evidence of a child penalty for [either] male [or] female applicants and no gender difference in the pay-off from a strong CV. (p. 407)

Experimental studies of hiring for non-tenure-track jobs. Although the above synthesis focuses on tenure-track hiring, there is a large research corpus that fell outside our inclusionary criteria because it concerned experimental and nonexperimental studies of non-tenure-track hiring. Hundreds of experiments, quasi-experiments, field audits, and observational studies of non-tenure-track hiring have been conducted, focusing on hiring of lab managers, civil servants, software engineers, postdocs, and student employees. In contrast to tenure-track hiring studies, these studies have often reported significant pro-male hiring advantages and double standards (e.g., see Koch et al.'s, 2015, meta-analysis; see also Foschi, 1996; Foschi et al., 1994; Reuben et al., 2014, Moss-Racusin et al., 2012). Some of these studies are highly influential and have been cited to support beliefs about bias in tenure-track hiring, even though

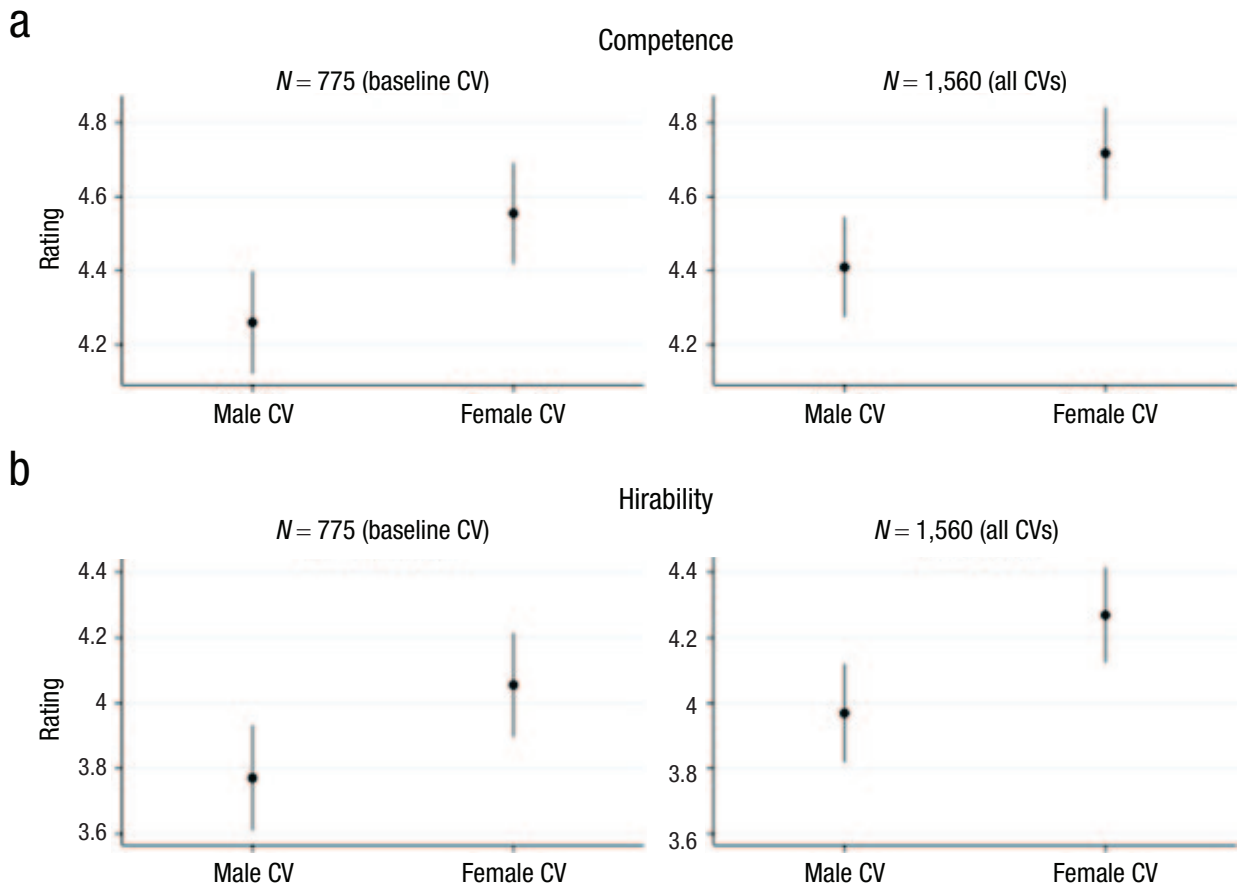

Fig. 4. Mean ratings of curricula vitae (CVs) for (a) competence and (b) hirability of male and female candidates, separately for analyses including only baseline CVs (left column) and all CVs (right column). In all graphs, the gender difference is statistically significant. Error bars show confidence intervals. Graph reproduced from M. Carlsson et al. (2021, Fig. 2).

none of them involved tenure-track hiring. This is conceptually problematic for reasons we offer below.

Koch et al.'s (2015) meta-analysis of 136 experimental studies of hypothetical hiring situations (not tenure track) explains why tenure-track hiring is unlikely to be gender biased. She and her colleagues found that bias was drastically reduced or completely absent when experienced professionals with motivation to make careful decisions chose whom to hire and when information about applicants was available that clearly indicated their high competence. In such cases, gender bias was nonexistent (.02 *SD*). Tenure-track hiring checks all three of Koch et al.'s boxes: Faculty doing tenure-track hiring are experienced professionals, applicants for tenure-track posts have high observable competence (CVs, talks, interviews),[7] and faculty are motivated to make careful decisions because unlike in typical experiments, their decisions result in lifetime colleagues.

Summary. The vast majority of findings—from (a) synthetic cohort analysis, (b) institutional hiring records, and (c) experiments—indicate that women are less likely than men to apply for tenure-track jobs, but when they do apply, they receive offers at an equal or higher rate than men do. Even though these three sources of evidence cannot be meta-analyzed, their findings, and those in powerful new experiments,[8] point in the same direction and are not consistent with claims of widespread bias against hiring women for tenure-track jobs. These conclusions extended to studies discussed here from the 1990s. However, there are no experimental studies of academic hiring between 1960 and 1990. Evidence from outside academia, including Schaerer et al.'s (2022) meta-analysis, suggests decreasing gender bias in hiring from 1976 to 2009; similarly, Birkelund et al.'s (2022) harmonized, cross-national callback analysis shows no discrimination against women in six countries differing along institutional, cultural, and economic dimensions.[9]

None of this means that women do not face very real barriers in completing their doctoral and postdoctoral training and segueing to tenure-track careers. For example, women are more likely than men to give up their initial aspirations to become tenure-track professors

while in graduate school, a finding primarily true of women with children or contemplating children. Undoubtedly, broad systemic factors are partly responsible, along with biological factors, for these women not applying for tenure-track positions. But the data do show that the reason women do not occupy a larger fraction of tenure-track positions is not because of a discriminatory tenure-track hiring process, as many researchers have alleged.

A counter to this conclusion is that in LPS fields (the very fields in which women are very well represented), female PhDs are less likely than male PhDs to apply for tenure-track positions, and this appears to be the primary reason why there are not even more female tenure-track assistant professorships in LPS fields than in the feeder PhD cohorts—something easy to overlook in view of the gender parity or even female superiority among assistant professors in LPS fields.

As shown in the large national analyses of actual tenure-track hiring by various universities and the NRC (2010) panel, female applicants in GEMP fields are usually either equally or more likely than men to be offered tenure-track jobs. Nearly all analyses in the past two decades accord with this conclusion, including the largest and best-controlled ones. Identical-CV quasi-experiments, which two decades ago revealed significant bias against women (Steinpreis et al., 1999), today do not. In the 22 years since Steinpreis et al.'s (1999) study, their finding has been supplanted by neutral and pro-female hiring results from larger studies (Breda & Hillion, 2016; Carey et al., 2020; M. Carlsson et al., 2021; Henningsen et al., 2021; Williams & Ceci, 2015). Identical-CV experiments for non-tenure-track jobs, particularly when candidates do not have unambiguously high qualifications (Eaton et al., 2020; Moss-Racusin et al., 2012) and when evaluators are neither experienced professionals nor have much at stake, are not predictive of the outcomes of hiring of excellent male and female tenure-track applicants, despite these non-tenure-track-hiring studies being cited more often than the latter (evidence presented in the Discussion section).

Evaluation Context 2: grant funding

In this section, we address the frequent claim of bias in grant reviews: "Understanding and targeting potential sources of bias in grant selection processes could be particularly important in improving the career advancement of women" (Alvarez et al., 2019, p. E9).

Grants are crucial to most of science. This makes the question especially important of whether funding agencies are more likely to fund grant applications from men than from women. We synthesize that literature here.

The obvious way to address this question is to compare men's and women's success rates: the probability that a grant application is funded. However, that is not the only way to pose the question. Whether a grant application is funded depends on the agency's assessment of the success of the research, and many researchers believe that PIs who have been successful in past research and publications are also likely to be successful in their currently proposed research. In fact, many grant agencies instruct reviewers to consider the publication record of PIs. And on average, men have more of a track record both in terms of publications and in terms of past grants. There are two reasons for this. First, in most fields and years, male PIs are older and therefore have a larger corpus of publications and are more likely to have been previously funded: Male full professors are, on average, 1.78 to 3.5 years older than women (e.g., van den Besselaar & Sandström, 2017; Brower & James, 2020), and in an analysis of 3,033 tenured and tenure-earning faculty members from 17 R1 universities in the United States, Samaniego and colleagues (2023) found that men were on average 3.8 to 5.50 years older. The second reason is that female scientists publish less than men in the same cohort and field, often because of more career interruptions resulting from family leaves (see Context 7). There have been several reviews of the grants literature (e.g., Ceci, 2018; Ceci et al., 2014; Ceci & Williams, 2011). There is one meta-analysis of many older studies (Bornmann et al., 2007) and a second meta-analysis of the same data 2 years later (Marsh et al., 2009). Here, we briefly review the pre-2006 studies and meta-analyses. We then describe more recent studies in more detail, including a meta-analysis that we ourselves have done, and provide dissectional analysis.

Grants before 2006. Many early studies have examined funding agencies outside the United States. In a highly cited article (1,981 Google Scholar cites as of May 24, 2022), Wennerås and Wold (1997) found Sweden's Medical Research Council (MRC) postdoctoral fellowships to be biased against women, even after controlling for research productivity. This study has been repeatedly invoked as prima facie evidence of gender bias in grants even after controlling for productivity, for example by Kaatz et al. (2014): "Lending support to this is the classic study by Wenneras [sic] and Wold in which female applicants for a postdoctoral research fellowship needed more than twice as many publications to receive the same competence scores as comparable male applicants" (p. 372). Notwithstanding this claim, Wennerås and Wold's methodology was problematic (see, e.g., Ceci & Williams, 2011, supplemental text S4; Hansson, 2009), and their data were lost (according to Wennerås and Wold), precluding reanalysis. Sandström and Hällsten

(2008) also analyzed Swedish MRC postdoctoral fellowships, now for 2004, using better statistical methodology than Wennerås and Wold's. In contrast to the latter's claim of antifemale bias, Sandström and Hällsten found a 10% advantage for women. They (and Wold subsequently, in Wold & Chrapkowska, 2004) note that institutional changes at the MRC over this period may have improved women's success rates.

There were many other studies of grants during this early period—some found bias against women, a few found bias against men, and many found small and insignificant gender differences (for reviews, see Ceci & Williams, 2011; Grant et al., 1997). In their meta-analysis of these 21 older studies, Bornmann et al. (2007) found that the probability that women were funded relative to men (odds ratio) was on average 7% lower for women, with variation ranging from 22% higher for women to 23% higher for men. However, they (along with colleagues) conducted a second meta-analysis of the same data 2 years later (Marsh et al., 2009), employing an improved methodology, and found "no evidence for any gender effects in favor of men, and even some evidence of an effect in favor of women. . . . This lack of gender difference for grant proposals is very robust" (p. 1311). For postdoctoral fellowships, they did find "a small, but highly statistically significant difference in favor of men," but concluded that "the size of this effect is sufficiently small that we still interpret [gender differences] as supporting a gender similarity hypothesis" (p. 1311).

We will not separately discuss all 21 of the early studies in the Bornmann et al. (2007) metastudy, except to note that only two of them controlled for productivity (one of which was Wennerås & Wold, 1997, discussed above), but we will summarize the few pre-2006 studies of U.S. granting agencies not included there. A large study of funding cycles from 1997 to 2004 (Hosek et al., 2005) for several U.S. agencies was not included in the Bornmann et al. (2007) and Marsh et al. (2009) meta-analyses. Hosek et al. (2005) found that women had lower success rates and received less money per award than men at the NIH (for 2001–2003 grant cycles), although an earlier study of the NIH found no such gender difference. At the NSF, Hosek et al. found wide differences year to year, ranging from advantages for women to advantages for men, with no time trend and no overall gender difference. A final study of earlier grants found that women and men at Harvard Medical School from 2001 to 2003 had similar grant success, "controlling for academic rank, grant success rates were not significantly different between women and men" (Waisbren et al., 2008, p. 207) and that there were also similar ratios in the proportion of money awarded to money requested.

As noted, only one pre-2006 study beside Wennerås and Wold (1997) had controls for research productivity: Bornmann and Daniel's (2005) study of German Boehringer Ingelheim Fonds for postgraduate fellowships from 1985 to 2000 found that when productivity was controlled, there was no gender bias.

In sum, pre-2006 evidence suggests that although some agencies evaluated men and women differently, on average they did not. Moreover, even if the flawed Wennerås and Wold (1997) research correctly identified bias, a later study by the same agency (Sandström & Hällsten, 2008) reversed their finding, showing that no bias existed by 2004, and even provided some evidence that women's grant applications at the same agency were favored.

Later grant studies and a new meta-analysis.[10] Because the only large meta-analysis of grant studies used data from before 2006 (90% before 2001) and analyses of grant awards have proliferated greatly since then, we conducted a series of meta-analyses of gender differences in grant awards starting with the award year of 2000 and ending in 2020 (see Kahn et al., 2022, which contains PRISMA diagrams, funnel graphs, and forest plots). We briefly summarize our approach and results here.

In order to have a truly comparable measure of grant success across all studies in our meta-analysis, we focused on one outcome measured in the same way: the gender differences in the percentage of applications that are funded. Of course, equal average success rates would not necessarily be evidence of gender fairness, for two reasons. First, one gender may have better research proposals or better past productivity. However, on the basis of past productivity, we would expect men to have higher success rates, so equal success rates are likely to suggest no bias against women or even pro-female bias. Second, men and women may work in different fields, countries, et cetera. Therefore, in addition to our basic measurement of effect sizes—which we report below as Cohen's *d*s (which equaled Hedges's *g*s in this study to three significant digits), the difference measured in terms of standard deviations of the success rate (Cohen, 1988, Formula 2.2.6), we controlled for several different moderators—country, years, fields, and when possible, past productivity.

The criteria for inclusion were that the study must have (a) data on both the number of grants submitted and grants funded by gender and (b) been published between 2000 and November 2020 in English and been based on data from 2000 to 2020. We first searched the Web of Science[11] and then supplemented this by searching the references of articles found for others we had missed. Often studies cited their data sources to be

online funding agency websites, and when possible, we used these websites to add additional years of data. When published studies were based on the very same agencies and grants in the same years, we were careful not to double count them.

It is important to stress that the goal of our metastudy was not to investigate what published studies claimed and argued but instead to aggregate the data on applications and grant awards to measure actual gender differences in success rates of grant applications and how they differed across various moderators.

All in all, our meta-analysis included 39 studies with data on 2,051,485 grant applications and 481,485 grants awarded by 27 different granting agencies. When we simply added all grants awarded and all applications by gender, we found that the gender difference in the likelihood of acceptance was 1.1 ppt out of an average success rate of 23.5%. However, this assumes a one-time competition for grants. When we used meta-analysis that allowed for random effects by study and agency and calculated the gender effect size measured as Cohen's *d*, we found a male advantage of 0.027 of a standard deviation, exactly equal to the 1.1-ppt difference in success rates without meta-analysis.[12] There was a great deal of between-study heterogeneity in effect sizes (I^2 = 91.3). In particular, effect sizes depended strongly on the location of the granting agency: United States, Canada, Europe, and elsewhere. We also separately measured effect sizes by location. Within the United States (where 84% of all grant applications we analyzed originated), the point estimate of Cohen's *d* (+ 0.005) suggests that women actually had a tiny advantage, on average, although this was not significant (p = .5). In contrast, in Europe, the male advantage (d = 0.041) was equivalent to 1.7-ppt difference in acceptance rates, whereas in Canada, the male advantage (d = 0.102) was equivalent to a large, 4.2-ppt difference in acceptance rates.

We also use multivariate methods to calculate results controlling for location, broad field, average year, and career stage of the grant program simultaneously.[13] With these controls, we found a time trend, where on average, the male advantage (Cohen's *d*) shrinks and the female advantage rises by 0.2 each year, equivalent to a decrease of male advantage of 2.4 ppt over the two decades (controlling for such variables as location). This means that even in locations with large average male advantages, these are considerably smaller by the end of the 20-year period.

We also found that in social sciences, controlling for other factors, the male-advantage effect size (Cohen's *d*) was lower (giving more advantage to women) than in other fields and particularly (d = 0.08) lower than in biomedical or physical science, equivalent to a 3.4-ppt difference in acceptance rates. Grant programs for more experienced researchers had a greater female advantage (d = 0.025) than the average grant program (equivalent to a 1.1-ppt difference in acceptance rates).

Eighty-two percent of the applications to U.S. agencies in our meta-analysis data set were to either the NIH or the NSF. These included all NIH research project grants, of which the majority were R01 grants. However, there are two types of R01 grants: new ones (Type 1 R01s) and renewals of previous R01grants on the same research project (broadly defined; Type 2 R01s). For the NIH, the success rates available by gender in the *NIH Data Book* (NIH, 2020), reproduced here as Figure 5, show that even without controls for men's greater productivity, the success rates of new (Type 1) R01s (NIH's largest grants category) have been identical for men and women for more than 20 years.

However, Figure 5 also indicates that the success rates have been higher for men for Type 2 R01s (renewals) every year except for in 2003, 2015, and 2019. This may be because men had more publications or patents from their original Type 1 R01 grants to justify these higher rates, given that Type 2 R01s are expressly dependent on the outcomes (publications, patents, etc.) of the original grants. Ideally, we would separate R01s by type in our meta-analysis and for Type 2 would be able to control for research products from the original Type 1 grant. These data are not publicly available. Thus, we cannot know whether the gender difference in NIH Type 2 success rates simply reflected the Type 1 R01's outcomes.

Type 1 R01s, the major NIH program for new awards, has been studied more than other NIH grants. Gender success rates for these have been equal since 2003, as is evident in Figure 5 and shown by various published studies. Ginther et al. (2016) is the only study of Type 1 R01s that controlled for productivity. It found that without controls, men and women had equal success rates in Type 1 R01s (2006–2010; with numerically lower rates among MDs). However, when controls for productivity were included, women were more likely to receive grants. Among experienced researchers, with publication and other controls, women and men were equally successful (in Type 1 grants; these data did not include Type 2 resubmissions). Women were also equally likely to be a finalist for R01s (receiving good priority scores). This suggests that, if anything, the NIH has a pro-female bias. (This same article separated out men and women by race and found that applications by African American researchers to the NIH are considerably less likely to be funded than those by White researchers.)

The NSF is another large U.S. government granting agency that has also been widely studied. On average,

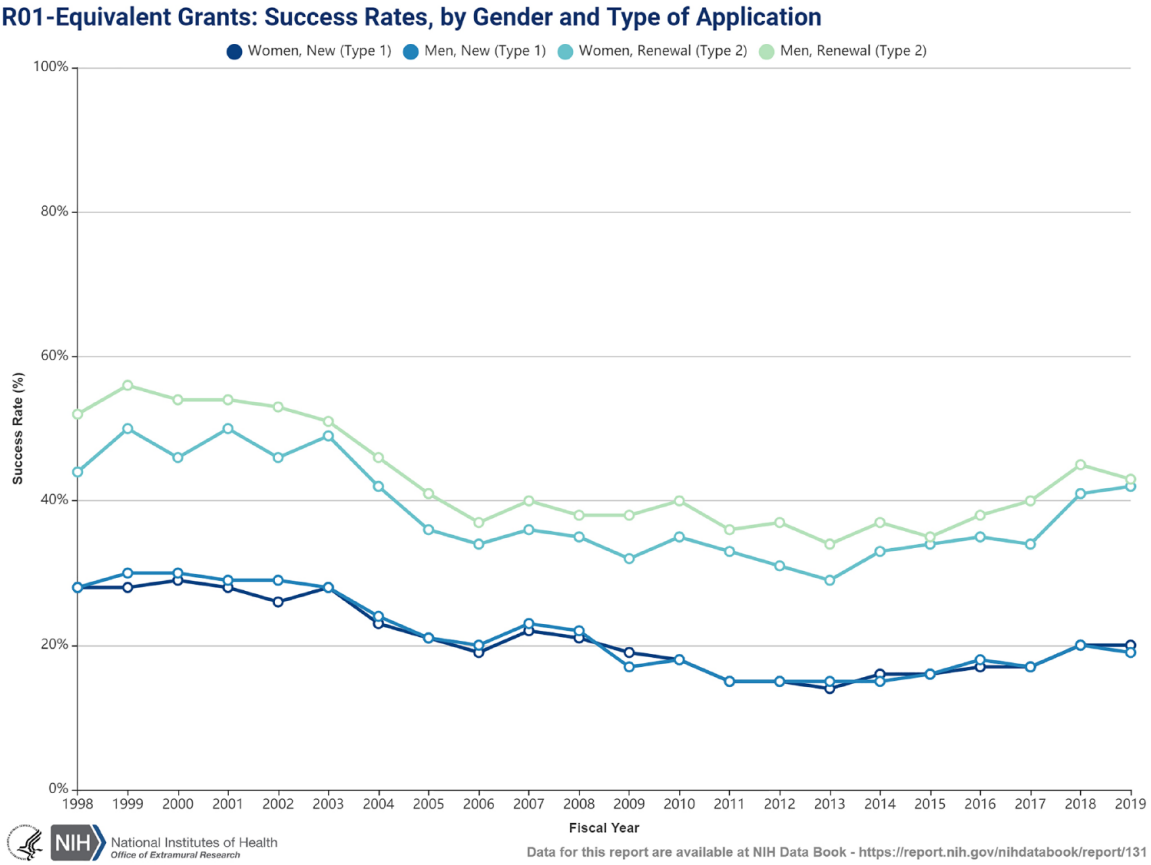

Fig. 5. Success rates between 1998 and 2019 for Type 1 and Type 2 R01 grants, separately for women and men. Graph reproduced from the *NIH Data Book* (National Institutes of Health, 2020).

women had a higher success rate than men at the NSF every year from 2000 to 2020. Over the 20 years, on average, the likelihood of grant success was 0.015 ppt higher for women than men (.269 vs. .254). Using the same data source but dividing it more finely, a large-scale analysis by Rissler and her colleagues (2020) of 15 years of grants across all six NSF directorates found similar results: Women were as likely to be funded as men, and there was no evidence of gender bias in any of the six areas of funding, with the exception of bias in favor of women in engineering when aggregated over the period from 2001 to 2016.

Rissler et al. (2020) also reported that the U.S. Government Accountability Office (2015) found no gender differences in success rates at three federal agencies: the NSF, the NIH, and the Department of Agriculture; at the Department of Defense, the Department of Education, and the National Aeronautics and Space Administration (NASA), there were either insufficient data to determine gender differences or evidence of disparities. Rissler et al. conclude as follows:

> Although our data suggest that women maintain equal success at receiving NSF research funding as men . . . we also show that fewer women submit research grant proposals as a PI relative to their representation in academia, especially in fields with more women. (p. 817)

Turning to studies of gender differences outside the United States (including international agencies), we find that male advantages in grant success were highly significant. As noted above, our metastudy for Canadian agencies (based on Appel-Cresswell et al., 2019; Burns et al., 2019; Tamblyn et al., 2018; Urquhart-Cronish & Otto, 2019; Witteman et al., 2019) indicated a large male advantage (d = 0.102, $p < .001$, a 4.2-ppt difference in acceptance rates).

Among these articles, there are two studies of grants from the Canadian Institutes of Health Research (CIHR) in biomedicine. One is Tamblyn et al. (2018), which concluded that gender bias existed, not on the basis of the success rates alone but also by a regression with controls for productivity. We discuss this study below.

In a second study of the CIHR, Witteman and her colleagues (2019) compared men's and women's results from 2011 to 2016 for two kinds of grants, one with an explicit emphasis (75%) on PIs' qualifications (past accomplishments, leadership, publications) and the other with an explicit emphasis (75%) on the project rather than PIs' accomplishments. They did not have productivity controls, but they did have age data. They found that when reviewers were told to give 75% emphasis to PIs' publications, leadership, et cetera, women had a probability of success 4.0 ppt (25%) lower than men's; when the emphasis was on the project rather than its PI, the probability of success was only 0.9 ppt (6%) lower for women.

Other papers in our meta-analysis studied 14 different funding agencies in Europe. As noted above, there was a male advantage (Cohen's $d = 0.041$, $p < .001$)—smaller than in Canada but still substantial. Articles included in Marsh et al.'s (2009) metastudy had also found substantial bias in pre-2006 European agencies that was larger than for agencies in the United States. However, subsequent analyses of European agencies have revealed less bias in some agencies: Mutz et al. (2012) analyzed nearly 8,500 proposals submitted to the Austrian Science Fund between 1999 and 2009, concluding, "We found that the final decision was not associated with applicant's gender or with any correspondence between gender of applicants and reviewers" (p. 121). On the other hand, long-term postdoctorate fellowships by the European Molecular Biology Organization (EMBO) reported large differences favoring men (20%) in success rates in 2006, which remained even when all references to PI gender were removed from the application and supporting letters, suggesting that men's proposals were viewed as stronger by reviewers who were unaware of the PI's gender (Ledin et al., 2007). However, later annual reports of the EMBO (e.g., EMBO, 2019) showed that things became more equal over time, so that by 2017, there was no significant difference in success rates except for some programs and years when women had an advantage.

Another study that showed some gender differences in a European agency after 2005 was van der Lee and Ellemers's (2015a) investigation of early career awards by the Netherlands Organization for Scientific Research (NWO). The authors found that from 2010 to 2012, women's average success rate was lower than men's (~4 ppt out of an average near 40%); however, in a commentary, Albers (2015) argued that after controlling for gender differences in discipline, this difference is not significant (and actually tilts toward pro-female bias), an example of Simpson's paradox, a claim disputed by van der Lee and Ellemers (2015b). Again, more recent evidence directly from 2019 shows that in both the NWO Vici and Vidi award programs over the previous 5 years, men and women's success rates were essentially equal (with women 0.1 ppt higher; NWO, 2023a, 2023b). Severin et al. (2020) analyzed scores of 12,294 proposals to the Swiss National Science Foundation and found a male advantage ($d = 0.17$), but after multivariate adjustments, this was reduced ($d = 0.065$). Finally, Bautista-Puig et al. (2019) analyzed all applications submitted to the European Research Council between 2007 and 2016, a total of 65,778 (10.8% funded), and found that across the three categories of European Research Council grants, women's success rate was significantly lower than men's (9.41% vs. 11.39%, respectively).

Some of the largest female disadvantages occurred at the Space Telescope Science Institute, which funds use of the Hubble Space Telescope. (We included this as a non-U.S. agency in our meta-analyses because the decisions were truly international.) Two articles about this agency's grants also provided an example of how reviewers' awareness of their previous productivity helps men. Men's proposals to use NASA's Hubble Space Telescope between 2001 and 2012 were granted more often than were women's proposals—23% versus 19% (Reid, 2014). Dissecting the 2017 cycle, at the first stage during which 150 astronomers evaluated 1,100 to 1,200 proposals, there was no gender difference in ratings. However, at the second stage, when a committee with access to the PI's track record decided among the highest-rated proposals, men's proposals were more likely to be granted. Because of these findings, NASA implemented a double-blind review process that excluded all identifying material, including gender. The result was a slightly higher success rate for women than for men (8.7% vs. 8.0%; Strolger & Natarajan, 2019).

The role of previous publications and previous grant success in determining grant acceptance is complicated. On the one hand, a PI's past productivity provides important information about the likelihood of grant success and future productivity. Because women have fewer publications than men (discussed below as Context 7), an unbiased evaluator concerned only about a project's success should take productivity into account. However, as Witteman et al. (2019) and Strolger and Natarajan (2019) showed, an emphasis on, or mere awareness of, productivity can lead to lower grant success for women that disappears when emphasis is put on the quality of the research itself. In recognition of the potential role of previous publications in tilting grant evaluations toward more senior (and thus more male) applicants, many agencies have separate research funds for early career scientists. In addition, some agencies have modified their biographical section to exclude information on past accomplishments and/or productivity or have

asked reviewers to de-emphasize these. This is less defensible with renewal grants such as NIH Type 2 R01s.

However, this gender-productivity-differential issue leads to a related question of whether reviewers evaluate publication records of men and women differently. If they do, and assuming that men's and women's publications are equally predictive or unpredictive of later productivity, then this differential evaluation could harm women. We are not aware of any study that has demonstrated such differential evaluation, except for Tamblyn et al. (2018) on the Canadian CIHR from 2012 to 2014. Whereas most studies in our meta-analysis had no productivity controls, Tamblyn et al.'s main analysis included two controls for productivity (*h* index and past funding success). Their abstract emphasized that one of these controls, past funding success, had a larger positive impact on success rates for men than women, concluding "there is evidence of bias [against women] in peer review" (p. E489). However, a careful look at all of their complicated regression results—with numerous interaction terms between gender and other variables—reveals other findings that indicate the existence of bias against men (larger positive effect of *h* indexes for women than for men, indicating that women were rewarded for higher-cited work whereas men were not and, also, higher scores for women having no previous grants and an *h* index of 0). Consequently, their reported empirical results do not provide evidence that men are rewarded more for productivity. (Of course, one could argue that controlling for productivity is not always appropriate because women, relative to men, have had less opportunity to be productive, to be leaders, etc. because of a greater frequency of caregiving and family leaves—coupled with the biological demands of childbearing—all of which result in fewer career opportunities, which might be at least partially related to biases.)

The best way to evaluate gender bias in grants is with a true experiment in which different evaluators are given the same proposal with male versus female names on it. There is only one such experiment, also based on NIH Type 1 R01s. Forscher et al. (2019) did an audit-study experiment, manipulating PI names on 48 NIH grants and sending them to reviewers for initial R01 evaluations. Clearly, this approach controls for everything about the application, including past productivity. The authors found no significant gender differences in ratings, with only tiny numerical perturbations across a very large array of measures.

One analysis of gender differences in a very different kind of grant funding situation reported bias against women even after double blinding. Kolev et al. (2019) analyzed proposals to the Gates Foundation's Global Grand Challenges Exploration. This grant differs in many ways from the research agencies' grants that were included in our metastudy, particularly in that the proposals were only a single page in length, the reviewers were not experts, the reviewers were given so many proposals ($N = 100$) that they could spend little time on each, and the reviewers did not discuss their reviews with other reviewers and panelists. Under these conditions, women did not fare as well as men: Women's proposals were approximately 15% less likely than men's to receive a high score despite the double-blind review procedure. Further analysis showed that a significant reason for this was that men used more broad language (words that appear across topical areas), whereas women used more narrow words that are specific to topics. The reason that the researchers interpreted such a difference as bias even though the reviewers were not aware of the PI's gender is because the change in women's ex ante–ex post performance was actually higher than men's (i.e., women received more subsequent NIH funding and published in more top-decile journals), thus calling into question the validity of the reviewers' ratings. However, these Gates grants represent a very different kind of contest than the standard, careful evaluation of lengthy grant applications by content experts who discuss their reviews with other panel members.

Taken together, both the analytic dissection and our meta-analyses appear not to support the claim that the grant peer-review process has been rigged against women PIs during the past 20 years in the United States. This is particularly true when analyses controlled for PIs' research productivity, a finding that accords with our meta-analyses, and also with Forscher et al.'s (2019) powerful random-assignment experiment. Using different analytic methods, other researchers have come to similar conclusions (e.g., Dehdarirad et al., 2015). Evidence from outside the United States is more concerning, and we hope that more studies with productivity controls or experiments help clarify the extent of the gender differences.

If the evaluation of research grants by U.S. granting agencies (controlling for productivity) is not biased,[14] then women's lower rate of funding relative to their representation among PhD-level researchers could be due to two factors: (a) women's lower average research productivity, which we will examine in Context 7, and/or (b) the fact that women apply less often for funding, even after being initially funded but especially after being declined for funding. The lower application rate of women as a factor in their lower funding level has been documented repeatedly (e.g., Broder, 1993; Ginther et al., 2016; Hechtman et al., 2018; Hosek et al., 2005; Ley & Hamilton, 2008; Pohlhaus et al., 2011; Rissler et al., 2020; Rockey, 2014; Sakai & Lane, 1996; Waisbren et al., 2008):

> The one thing that is almost always true . . . is that fewer women submit proposals than men. (Rissler et al., 2020, p. 815)

> Future work looking beyond administrative data may address why women in academia might not be reapplying (or applying) for [research project grants] at the same rates as men. (Hechtman et al., 2018, p. 7947)

To conclude this section, it deserves reiterating that a failure to find evidence of gender bias in the United States—in this case, in the context of claims of biased grant reviewing—means only that and nothing more. Women may still face systemic barriers that impede them from submitting as many grant applications as their male counterparts (Rissler et al., 2020), such as greater familial responsibilities and biological demands of child-rearing. It would be informative if data existed on submission (and resubmission) rates of women with young children versus others.[15] This is, however, a different question than asking whether, once a grant is submitted, reviewers may be biased against women (the claim that is often made), but it is one worth pursuing.

Evaluation Context 3: teaching ratings

It is frequently claimed that students downrate female instructors:

> [Women] receive systematically lower teaching evaluations despite no differences in teaching effectiveness. (Witteman et al., 2019, p. 531)

> Research on teaching and learning has recognized and documented significant gender and racial biases with teaching evaluations completed by students. . . . Evaluations of white men are often higher than those of women and faculty of color . . . whose evaluations are artificially deflated due to biases. . . . Students bring gendered expectations into the classroom and evaluate professors according to whether the instructor fulfills or fails to meet these stereotypes. Given the predominance of men in academia, effective instruction has become synonymous with masculine characteristics. (Key & Ardoin, 2019, pp. 1–2, citations omitted)

Our synthesis below of numerous metastudies on this topic and the hundreds of studies they are based on support these claims by Witteman et al. (2019) and Key and Ardoin (2019). We find that female instructors appear to be downrated at least in some contexts—depending on factors such as academic discipline, gender of rater, gender mix of class, gender mix of faculty, gender match, and student's expected grade. Bias against women is especially evident if we take into account qualitative comments that some students attach to their end-of-course numerical ratings, comments that are more abusive toward women instructors (e.g., Heffernan, 2022).

We do not base our conclusions regarding gender bias in teaching ratings on a meta-analysis of our own. This is because there have already been several meta-analyses of bias in teaching ratings, including some that provided narrative assessments of this large literature. The conclusions of these metastudies differed, with some authors concluding that gender bias against female instructors exists, and others concluding that teaching evaluations are gender neutral. Here, we provide a "review of reviews" of this literature, in which we undertake an analytical dissection of the bases of the differing conclusions produced by these sometimes contradictory analyses. (Two studies that were published after we completed our analysis support our conclusion of gender bias in teaching evaluation—Buser et al., 2022, and Chatman et al., 2022.)

Importance of moderators. At the outset, we note that myriad moderators have been documented, including academic discipline, time of class, instructor's attractiveness, student's expected grade (although less so in higher-ranked universities), grade distribution within the overall class, grade inflation, size of class, level of class, gender of student, proportion of female students, gender match between instructor and student, ethnicity of student, language background of instructor, age of instructor, interaction between student gender and rank of instructor (PhD student, lecturer, professor), interaction between instructor gender and level of course, and even the number of scale points on the rating instrument[16] (e.g., Bachen et al., 1999; Centra & Gaubatz, 2000; Doubleday & Lee, 2016; Hamermesh & Parker, 2005; McPherson et al., 2009; Mengel et al., 2019; Rivera & Tilcsik, 2019).

As noted above, bias against women is particularly apparent if we take into account qualitative comments that some students attach to their numerical ratings, comments that are more abusive toward women instructors (e.g., Heffernan, 2022), especially toward those from non-English-speaking backgrounds (Fan et al., 2019). For example, Schmidt (2020) scraped over 14 million ratings on the Rate My Professors (RMP) website and found that negative terms were used more often for female instructors, whereas some positive terms were used more often for male instructors.[17]

In our synthesis, we focus on research about student evaluations of courses taught rather than on less relevant research on student evaluations of single videos,

single stand-alone lectures, and the like (e.g., Abel & Meltzer, 2007; Basow et al., 2013; Graves et al., 2017). We describe our evidence below. When possible, we have converted findings into a common metric that we calculated (Cohen's *d*—the gender effect size expressed as the proportion of a standard deviation) to aid in comparisons, unless the study does not include sufficient data for this calculation. We start by briefly describing results of the meta-analyses that exist and why they differ.

There are 11 metastudies of teaching bias, but only four focus on gender: Feldman (2007), Kreitzer and Sweet-Cushman (2022), Heffernan (2022), and Wright and Jenkins-Guarnieri (2012). The first two conclude there is no gender bias, the last two that there is. We first discuss the two that found no gender bias.

The influential review by Feldman (2007; 663 Google Scholar cites) argued,

> A recent review . . . of three dozen or so studies showed that a majority . . . found male and female college teachers not to differ in the global ratings they receive from their students. In those studies in which statistically significant differences were found, more of them favored women than men. However, across all studies, the average association between gender and overall evaluation of the teacher, while favoring women, is so small (average $r = +.02$) as to be insignificant in practical terms. (p. 97)

The recent review he mentions was his own 1993 study. The other meta-analysis that found no bias was that of Wright and Jenkins-Guarnieri (2012). However, Wright and Jenkins-Guarnieri included only one meta-analysis related to instructor gender: Feldman (1993). In other words, the only metastudy that found no gender bias was the 30-year-old Feldman (1993) article, reviewing studies from 1979 to 1991.

Below, we summarize and dissect the more recent meta-analyses by Kreitzer and Sweet-Cushman (2022), which included more than 1,000 prior studies, and Heffernan (2022), which covered 136 studies. We focus on some of the key individual studies in those analyses. Both of these metastudies led to the same conclusion we have reached below: Namely, that although there is some evidence of gender neutrality in many studies of student evaluations of faculty, when all forms of data (numerical, qualitative, experimental, correlational) are taken into account, there is some evidence of bias against female instructors. However, the contexts are extremely important.

Qualified evidence of gender bias. The second largest study showing correlational evidence of bias against female instructors is Rosen's (2018) analysis of nearly 8 million RMP ratings (in which teacher gender was assigned by an algorithm). Although RMP ratings are samples of convenience not designed to be representative, the same can be said of standard institutional course evaluations (Feldman, 1993; Rosen, 2018), and we find that both types lead us to similar conclusions. RMP is the largest corpus of students' ratings and contains 20 tags such as "tough grader" that allows unique tests along many dimensions.

Rosen (2018) found that, across all subjects, men on average had a small advantage in students' rating of overall quality (mean Cohen's $d = \sim 0.10$). However, as can be seen in Rosen's Table S1 (in their supplemental data), for some aspects of the rating in some disciplines, the differences were huge, such as a male advantage for "clarity" in history (Cohen's $d = 0.35$). Wallisch and Cachia (2019) analyzed an even larger corpus of RMP evaluations and also found significant correlations between ratings and gender; again, male instructors received a mean rating that was higher than females (by 0.046 on a scale from 1 to 5).

Gender differences in ratings also seem to differ by the gender of the student and the gender mix of the students. Men are usually but not always more positive toward male teachers. For instance, in economics classes at a state university, Mengel et al. (2019) studied the gender difference for approximately 1,000 evaluations. Male students evaluated female instructors as being worse ($d = 0.207$) than male instructors, whereas female students evaluated female instructors better than male students did but still as worse than male instructors ($d = 0.076$). Fan et al. (2019) also found that male students favored male instructors, controlling for language of the faculty and whether raters were international students. Boring (2017) found that although both female and male students downgraded female faculty, male students did so to a much greater extent. Using observational data, Funk et al. (2019) found that female students rated female faculty higher than did male students (and they rated male faculty higher than did male students as well), but only in subjects with low shares of female faculty. Earlier, Bachen et al. (1999) found that female students rated female faculty higher than male faculty but that male students rated both the same.

Gender differences in ratings also depend on the subject, although there is massive disagreement about which subjects are associated with women being penalized. For more mathematical courses and/or male-dominated courses, the results from various studies are contradictory.

In Rosen's (2018) correlational study, quantitative courses revealed insignificant gender differences; for example, female instructors had the advantage in math (Cohen's $d = 0.04$) and chemistry (Cohen's $d = 0.004$), whereas male instructors were rated better in psychology (Cohen's $d = 0.03$). Price et al. (2017) studied teaching evaluations at a Swedish engineering school where computer science was mostly taught by male professors and approximately half of environmental engineering instructors were women. In the former, there were insignificant gender differences in evaluations of "good teaching" and "high generic skills." But in the latter, men got significantly higher ratings, particularly for "good teaching."

However, other studies have reported the opposite. Rivera and Tilcsik (2019) found for a large U.S. university (with 100,000 evaluations) that in male-dominated subjects, gender differences were large when a 10-point rating scale was employed: In the least male-dominated fields, gender differences in evaluations were insignificant, but the gap was related to the number of scale points available to students. When a 10-point scale was used, the same instructor received a mean rating of 7.8 when students thought the person was a man but a mean rating of 7.1 when students perceived the instructor to be a woman ($p < .05$). When the 6-point scale was used, the gap was not significant: a mean rating of 4.9 ($SD = 0.9$) when students perceived the instructor to be male versus a mean rating of 4.8 when they perceived the instructor to be female. Similarly, Mengel et al. (2019) found that with random assignment of nearly 20,000 Dutch students, male faculty's advantages were largest for courses with strong mathematical content when rated by either men or women: $d = 0.32$ when male students rated, $d = 0.28$ when female students rated. For courses without math, male advantages were smaller: $d = 0.17$ for male raters, $d = 0.04$ for female raters.

Thus, some studies have found that the greatest penalties for women are in less mathematical subjects, including humanities and business, whereas others have observed the biggest penalties in science, engineering, and math classes. Although further research is needed to reconcile this issue, we conclude, as did Kreitzer and Sweet-Cushman (2022) in their large metastudy, that the preponderance of data suggest that female instructors fare better in humanities than in natural and social science, although there are divergences from this conclusion.

Note that gender differences in teaching evaluations, even with controls for such variables as content, may underestimate bias if students can choose their courses and professors, given that students may avoid faculty they would not like. This potential confound is precluded in studies of identical required courses for which students are randomly assigned to professors. Most of the studies with random assignment were conducted in Europe. Boring (2017) compared professors at a French university in six mandatory social science courses. As noted above, she found that both male and female students rated female instructors lower on overall satisfaction. Similarly, Mengel et al. (2019) found that when Dutch students were randomly assigned to instructors in multisection business and economics courses, female instructors received lower evaluations from both female and male students, $d = \sim 0.20$ by our calculation (using approximately 20,000 evaluations). However, the authors found that instructor gender did not affect grades or study hours.

A third random-assignment study (Wagner et al., 2016) analyzed 688 evaluations of two-person faculty teaching teams—some same gender, some mixed gender—in social studies courses at a Dutch university with international master's students. In their experimental design, self-selection into courses but not instructors was allowed. Applying various controls, the authors found that women received lower evaluations (which we calculated as $d = \sim 0.28$) for identical courses and years. Finally, in the United States, Mitchell and Martin (2018) found a gender difference of $d = \sim 0.38$ in 1,090 ratings for a randomly assigned introductory political science course for which most content was online, although there was some contact via office hours and email with two actual professors (similar ages), one male and one female.

The most unequivocal experimental evidence of bias comes from MacNell and colleagues (2015), albeit the evidence is limited by the study's small sample size. Students taking an online anthropology/sociology course were told their instructor's name was Paula or Paul, by random assignment. Regardless of what they were told, they randomly had a woman or a man as an instructor. (The latter randomization was necessary to eliminate the possibility of students reacting to gender differences in the tone, content, or language of online communication.) Although this study had only 43 students, the instructor received significantly higher ratings when called Paul, especially from female students. In contrast, the differences by actual gender were insignificant. As one journalist noted, "The results were astonishing. Students gave professors they thought were male much higher evaluations across the board than they did professors they thought were female, regardless of what gender the professors actually were" (Marcotte, 2014, para. 3). We estimate the effect size of the gender difference in this study at $d = \sim 0.50$ of a standard deviation. Granted, it was a very small experiment, but it is nevertheless suggestive, given that it contained the most stringent experimental controls in the entire literature.[18]

It is important to point out that these discrepancies cannot readily be explained by differences in instructor quality or students' investment of study time. Although students rate women instructors lower, they do not learn less if they are taught by women (as observed by Mengel et al., 2019, and Linek et al., 2009). (Other studies, including meta-analyses, e.g., Uttl et al., 2017, also found that learning did not depend on instructor gender.) Gruber and her colleagues (2021) provide a synopsis of this issue: "[Grades] and student learning are weakly correlated with student ratings in both experimental studies [and] real-world teaching contexts" (p. 495). Boring (2017) also found that "students appear to learn as much from women as from men" despite their evaluation differences (p. 27).

Bias in language terms. Some studies analyzed themes, salutations, and adjectives, or tags in RMP that students used to describe professors (e.g., "hot"). Mitchell and Martin (2018) found that students were more likely to refer to male instructors as "professors" and to female instructors as "teachers." Schmidt (2020) provides an interactive platform that allows users to see how any word is used to describe female versus male faculty in 14 million RMP reviews. Using it, we found a similar gendered gap for "professors" versus "teachers." However, Key and Ardoin (2019) found a more complicated result, with men referred to as "teacher/instructor" more than women but only slightly more often referred to as "professor." C. C. Miller (2015) found that in RMP, students more often referred to female instructors as "bossy" and male instructors as "assertive." Words such as "genius" were used more often to describe male instructors. Again using Schmidt's platform, we found that nearly all negative words we entered ("unkind," "cruel," "bossy," "mean," "disorganized," etc.) were more frequent in comments made about female instructors.

Wallisch and Cachia (2019) performed regression analyses of more than 800,000 RMP professor profiles to determine whether gender could be predicted by tag differences (e.g., "tough grader," "amazing lecturer"). They found little evidence in support of substantial gender differences in the mean rating of teaching evaluations for in-person classes; they found similar results in a follow-up study of online courses, massive open online courses, et cetera. Other researchers have also analyzed written comments that students add to their numerical ratings and have also found gender differences favoring male instructors (Key & Ardoin, 2019; J. Miller & Chamberlin, 2000; Schmidt, 2020; Storage et al., 2016). In a recent review of 183 articles on student evaluations, Heffernan (2022) categorized the themes in students' comments, finding bias against women and minorities: He found that gender—even perceived gender—results in student evaluations that are highly prejudiced against women (see also Rivera & Tilcsik, 2019).

Teaching evaluations are often part of tenure and promotion deliberations (e.g., Jones et al., 2014; Linse, 2017; McPherson et al., 2009; Mengel et al., 2019; Uttl & Smibert, 2017, 2021). McPherson et al. (2009) state that evaluations of faculty teaching "are commonly used to inform decisions about merit-based raises and are often an important component in the promotion and tenure process" (p. 48). Seldin (1993) reported that 86% of U.S. faculty evaluations employ student evaluations of teaching as a major criterion. And Wagner et al. (2016) reported that at their Dutch university, "women are 11 percentage points less likely to attain the teaching evaluation cut-off for promotion to associate professor compared to men" (p. 79). Clearly, the possibility of gender bias in ratings argues against their use for high-stakes personnel decisions.

On the basis of their analysis of more than 100 studies, Kreitzer and Sweet-Cushman (2022) criticized past meta-analyses for various reasons, such as omission of qualitative linguistic comments that some students make in addition to their numerical ratings. Our own analysis accords with their conclusion, with the proviso that if one were to downplay the RMP data because they are a convenience sample of unknown representativeness, then the evidence for gender bias becomes much weaker:

> While a few meta-studies find few gender differences (Wallisch & Cachia, 2019; Wright & Jenkins-Guarnieri, 2012), the vast body of literature across time and methodological approach consistently finds the opposite. Research has demonstrated a multitude of ways that men benefit from evaluation, while women do not fare as positively. (Kreitzer & Sweet-Cushman, 2022, p. 76)

In sum, the evidence supports the claim that female instructors are penalized for being women, independent of the content and delivery of their lectures and independent of students' actual learning. The effect sizes we calculated indicate penalties for women that ranged between small and moderately large (ds = 0.10–0.50). So, unlike the domains in which we were able to unequivocally reject claims of widespread gender bias, in this domain, we conclude that there is gender bias. However, we supplement this conclusion with Linse's (2017) caveat that gender biases in student teaching evaluations

> definitely exist [but] rarely, if ever, fully explain the student ratings results. . . . Over time, a growing body of research has been able to document gender effects on student ratings, but these effects are

> neither uniform nor consistent across all disciplines, nor do they apply to all women. (p. 98)

This caveat accords with our interpretation of the contextual nature of the findings. Finally, we emphasize that instructor evaluations cannot be explained by differences in instructor quality or student learning; the best evidence indicates a general lack of correlation between evaluations and student learning outcomes (e.g., Uttl, 2021; Uttl et al., 2017).

Evaluation Context 4: journal acceptances

A frequent claim is that women's journal and conference submissions are held to a higher standard than men's (e.g., Budden et al., 2008; Ferber & Teiman, 1980; Knobloch-Westerwick et al., 2013; Lortie et al., 2007; Murray et al., 2019; Roberts & Verhoef, 2016; Walker et al., 2015): "Research on anonymous refereeing shows pretty clearly that biases play a role in evaluating work" (Saul, 2009, para. 4).

If women's publications are more likely to be rejected, this could explain why they publish less than men (see Context 7) and perhaps why they receive less R01 Type 2 funding than men. A first pass at testing this hypothesis would be to evaluate whether the acceptance rate is lower for women's articles using journals' administrative data on numbers of submissions and acceptance rates. To systematize the studies that have addressed the question of gender differences in journal acceptance rates, we undertook a formal meta-analysis of this literature (Kahn et al., 2022). Here, we summarize that meta-analysis and then discuss the issues raised in this literature.

In our meta-analysis, we analyzed gender differences in acceptance rates of manuscripts submitted to scientific journals over the first two decades of this century, from 2000 to 2020. We included all studies that gave data on numbers and success rates of submissions (rather than those that merely described gender breakdowns of articles published without data on acceptance rates). While we looked for data about final acceptance rates, we also included studies where we knew about only one stage of the review process (e.g., Did it get a revise and resubmit [R&R] decision? What were the reviewers' recommendations?)

As we did for grants, we started with a systematic search on the Web of Science. We supplemented this by searching through each of these article's citations for more possible articles. Our prespecified inclusion criteria captured 33 articles from which we were able to derive 79 mutually exclusive subsamples of submissions. These 79 base cases represented 410,504 journal submissions across various fields of science. Studies differed in their definition of gender of multiauthored articles (first author, corresponding author, etc.) and by scientific field.[19] In some of our analyses, we controlled for this, as well as for year, field, whether the process was double blind, and whether the authors controlled for personal characteristics (either by regression or by running separate analyses).

Overall, the gender effect size, measured as Cohen's d (or Hedges's g), showed a small male advantage of 0.024 for studies that gave probabilities of acceptance or R&R (d = 0.028 including studies reporting other outcomes). To put the scale of this number into context, consider that for the 65 studies that gave probabilities of acceptance or R&R, the weighted overall average acceptance rate was 28.6%, and d = 0.024 equals a difference of only 0.005 in acceptance rates, a small difference relative to average acceptance rates (but statistically significant because of the large sample size).

There was considerable variation in effects sizes across studies (as measured by the heterogeneity index I^2), suggesting roles for moderators. Isolating studies by evaluator roles (reviewers or editors), we found some bias attributable to the reviewers (d = 0.051) but none to the editors. Separately estimating effect sizes by field, we found that fields differed, with economics having the largest gender differences (d = 0.096) and physical science the second largest.

We also investigated whether there were also differences across female definition (first, last, corresponding author) and whether there were time trends. Because female definition of the study was likely to be correlated with fields, we wanted to control for these. However, because author position is purely alphabetical in economics, we left it out of this analysis. Using multivariate techniques to control simultaneously for field, female definition (based on first, last, or corresponding author), and time (but excluding economics), we found that whether female was defined using first or last author did not make a difference but that those studies using corresponding author showed a greater male advantage (but p = .08). There was also a significant positive time trend toward more female advantage. The time trend was large enough that for any field except economics and when female was defined as either first or last author, by 2020, the point estimate suggests that women are more likely to have their papers accepted but only significantly more likely for those in social sciences outside of economics.

Finally, some studies also had productivity moderators. In a separate analysis including only these studies, there was no male advantage, and instead women had an advantage in acceptances (because men have higher productivity, as shown in Context 7).

After we concluded these meta-analyses, a large team of data scientists (Squazzoni et al., 2021) reported the results of the biggest-ever study of gender bias in journal acceptance rates, based on over 740,000 submissions. In line with the findings from our meta-analyses, Squazzoni et al. also found no bias against female authors—defining *female* as first or last author or the proportion of female authors overall and within broad fields. In fact, they found a slight advantage for female authors overall. If their massive analysis had been published months earlier within our temporal inclusion window (2000–2020) and had thus been included in our meta-analysis, it would have added further weight to the claim of no systematic bias against female authors.

Thus, neither Squazzoni et al.'s (2021) study nor our meta-analysis support claims of clear pervasive gender bias in publication acceptance rates, particularly in recent years. Therefore, below, we provide a narrative review to supplement the meta-analysis because it addresses a number of conceptual issues and posits that averaging across studies can sometimes lead to conclusions that are misunderstood or alleged to prove things they do not in fact prove. This narrative dissection of studies measuring gender differences in publication decisions includes analyses of premier journals such as *Science* (Berg, 2017, 2019; Braisher et al., 2005), *Nature* (Braisher et al., 2005; McGillivray & De Ranieri, 2018), *PNAS*, and *Cortex* (Brooks & Della Sala, 2009; Valkonen & Brooks, 2011).

Inconsistent evidence of gender bias. Although there is no prima facie evidence of pervasive gender bias, there may be bias in specific fields, at specific times, and/or in specific journals. For instance, Berg (2017) analyzed acceptance rates of articles submitted to *Science*. He studied a sample of 2,650 accepted manuscripts in 2015 and a similarly sized not-accepted sample. He found no significant gender differences for either, with numerically higher acceptance rates for junior first-author women and lower rates for senior women. Berg and his team have since analyzed a larger set of 66,057 articles published in *Science* between 2010 and 2017 (Berg, 2019) but to date have published only results for the category termed "Reports." As in their 2015 sample, there were no significant gender differences in acceptances over the period from 2010 to 2017 for either "first author" or "corresponding author." Year to year and across fields, both genders' advantages in acceptance rates fluctuate. However, there are some differences by field. First, in physical sciences, acceptance rates for male corresponding authors were higher than for women for the period from 2012 to 2016 (although Berg found the same acceptance rates for women and men before and after this period), whereas acceptance rates were equal for male and female first authors since 2013. The second difference is that in the life sciences, women's acceptance rates were higher than men's from 2016 to 2017 for both types of author.

As another example, Murray et al. (2019) studied approximately 30,000 submissions to the biosciences journal *eLife* from 2012 to 2017. Their results showed a small but significant male advantage if a woman was the corresponding author (odds ratio = 1.057, $p = .047$) or if a woman was the last author, but no significant difference if a woman was the first author ($p = .56$).

Card et al. (2020) argued that although there was no gender difference between acceptance rates in their study, even after controlling for factors such as numbers of past publications, this might not ensure that the quality of men's and women's accepted articles are similar. Instead, they argued that only subsequent citations to the accepted articles signal quality. Analyzing citations, Card et al. found that in economics, accepted women's articles had higher subsequent citations, and from this they concluded that bias against women exists, despite no gender differences in acceptance rates. There might indeed be some gender bias in economics publication evaluation, particularly given that there is evidence from another study showing that the quality of writing in published articles in economics by women was higher than in articles by men, whereas the time until women's articles were finally accepted was longer, suggesting bias (Hengel, 2022).

However, many reviews across a broad swath of STEM fields have found no differences in citation rates per article by men and women (e.g., Ceci et al., 2014; Lynn et al., 2019) or even higher citations to papers written by male authors, a result not due to selective citing of male papers by male authors or gender homophily (for a review of evidence against gender citation homophily, see Tekles et al., 2022). By Card et al.'s (2020) logic, this suggests that there is no overall bias against women. In addition, many articles have documented lower citations per article for women in specific journals or fields (e.g., Maliniak et al., 2013; Odic & Wojcik, 2020). Again, using Card et al.'s logic, this suggests bias by the journal in favor of women. Maliniak et al. instead interpret their finding of lower citations for women as bias by the readers against women. Clearly, it is difficult to know what information can be inferred from gender differences in citations per article, because the same outcome (citations) has been interpreted oppositely by different parties and reviews have disagreed on whether there is citation bias.

Role of single-blind versus double-blind review. Scientists commonly acknowledge that blind review is the pinnacle of peer review:

> Wherever possible, reviews should be done blind, so the reviewer does not know whom they are reviewing. A well-known example of the effectiveness of this technique is in orchestra auditions, where the proportion of women hired shot up when auditions were performed anonymously behind a curtain. (Urry, 2015, p. 473)[20]

Because of difficulties in controlling for quality, some researchers have investigated double-blind reviewing (e.g., Bernard, 2018; Tung, 2006), meaning that neither authors nor reviewers know the identity of the other, or even triple-blind reviewing, where neither authors, editors, nor reviewers know each other's identity (in single-blind reviewing, the author is unaware of the identity of the reviewers, but the reviewers know the authors' names). If gender gaps in acceptance under nonblind review are narrowed by blind review, then this suggests bias in the evaluation of the quality of women's submissions.

The best way to compare double- with single-blind reviewing in journal acceptance is to randomly assign papers. The first study to do this was Blank's (1991) double-blind-reviewing experiment at the *American Economic Review*. She found that double-blind-reviewing did not significantly change women's acceptance rates relative to men's, either with no controls or after controlling for quality using the author's institutional prestige as a proxy (both ps > .7).

One issue with contrasting double- and single-blind reviewing is that reviewers might detect the identity of masked authors (e.g., recognize the work from conference presentations). In older studies, there is mixed evidence for this, with one study finding that reviewers for a half dozen mainstream psychology journals were unable to identify authors during blind review (Ceci & Peters, 1984). Blank (1991) found that about half of the reviewers did in fact know the author's identity and gender. However, even when she used the subsample of reviewers who did not detect the author's identity and gender, she still found an insignificant impact of double-blind-reviewing on the gender difference in acceptances (p = .38). (In these days of Google Scholar and preposts, double-blind-reviewing without reviewers being able to know the author is virtually impossible.) Blank (1991) acknowledges that "The lack of significance for the gender differences . . . reflects the small number of women's papers in the data set" (p. 1053).

The second randomized experiment was by Tomkins et al. (2017a), who focused on reviewers of articles submitted to be published in conference proceedings. Half of the program committee members were randomly assigned to double-blind review, in which they did not see the author's identity, and each of the 500 articles was assigned to two reviewers, one performing double-blind review and one not. Double blinding did not significantly change the review scores (p = .18), although point estimates suggest that women fared better with double-blind review.

The third randomized experiment was by F. Carlsson et al. (2012), in which half of the reviewers of 940 submissions for an economics conference were randomly assigned to double-blind review. The authors found no significant difference in the scores given to women in the nonblind versus blind sample (ps = .60–.99), nor did they find any significant gender difference in the average scores.

These few studies are the only randomized controlled trials of journal and conference acceptances that we know of. In the previous section, we described two randomly assigned double-blind experiments on grants (Kolev et al., 2019; Ledin et al., 2007), both of which found that gender differences in grant awards that occurred under nonblind review conditions remained even under double blinding of reviewers.

Other studies comparing double- with single-blind reviewing looked at differences in journal acceptance rates when a journal moved from one type of review to the other. In a highly cited article, Budden and colleagues (2008) reported that the acceptance rate for women increased 33% (9.3 ppt) for the journal *Behavioral Ecology* after it changed from single- to double-blind reviewing. However, before–after comparisons are problematic when they are not contrasted with similar before–after cases with no treatment change. (In economics terms, one needs to calculate the differences in differences to avoid potential confounds, and even then, pretrends may be inaccurately identified; Kahn-Lang & Lang, 2020.) After Budden et al.'s study was published, other researchers noted that ecology journals that did not change from single- to double-blind reviewing also witnessed an increase in women's acceptance rates during the same period, suggesting that double-blind reviewing had not caused the change (Engqvist & Frommen, 2008; Hammerschmidt et al., 2008; Webb et al., 2008; Whittaker, 2008).

The Royal Society of Chemistry (2019) found that moving from single- to double-blind reviewing significantly increased acceptance rates of women's papers relative to men's but by very small amounts (0.3 to 1.0 ppt). However, they, too, did not contrast this with similar journals that did not change policies.

Roberts and Verhoef (2016) studied the shift from single- to double-blind reviewing at Evolution of Language (EvoLang) published conferences. Averaging over 2 years of conferences (EvoLang 9 and EvoLang 10), they found no significant gender gap in single-blind reviewer scores for men and women submitters.

Then, after the shift to double-blind review in EvoLang 11, reviewer scores for women's papers were higher than for men's, from which Roberts and Verhoef concluded, "double-blind reviewing at EvoLang 11 reveals gender bias." However, examination of their results shows that from EvoLang 9 to EvoLang 10 (both of which employed single-blind reviewing), women's scores also increased relative to men's. The increase from single-blind review (EvoLang 10) to double-blind review (EvoLang 11) was not significantly different from the previous trend. Moreover, Cuskley et al. (2020) repeated Roberts and Verhoef's analysis for the subsequent EvoLang 12 conference—still double-blind review—and found no female advantage, again undermining both the claim of bias and of double-blind review being an improvement for women.

McGillivray and De Ranieri (2018) analyzed manuscript acceptance rates by 25 journals that are part of the *Nature* family (128,454 submissions over a 2-year period). They reported that the acceptance rates for both men and women were no different under single- and double-blind review. However, double-blind review had a cost: Manuscripts (both men's and women's) were sent out for review at a lower rate than under single-blind review.

Østby et al. (2013) also did a before–after comparison of single- to double-blind review in an international relations journal, finding no effect of moving to double-blind review (or any overall gender difference in acceptance, either before or after the adoption of double-blind review). Finally, Heath-Stout (2020) analyzed submissions to the *Journal of Field Archeology* between 2009 and 2013 (when it used single-blind review) and 2014 to 2018 (when it switched to double-blind review). Overall acceptance rates went down under double-blind review, but this did not differ by gender of author, and there was no significant gender difference in acceptance rates in either period.

Thus, the main lesson from studies of moving from single- to double-blind review is that the comparison cannot be used to identify bias if there are no appropriate counterfactuals to indicate what would have happened without this change.

A final approach to evaluating bias is experimental, randomly assigning the same article to reviewers but changing the gender of the authors' names. Such experiments have used only student or postdoc reviewers, who lack faculty's skills, knowledge, and time to fully evaluate the work. As we found earlier in the contrast between Koch et al.'s (2015) meta-analysis of experimental hiring studies using student evaluators versus faculty evaluators, students' evaluations are typically more gender biased than faculty's. Among these experiments with students, some found that male names were favored (Krawczyk & Smyk, 2016; Knobloch-Westerwick et al., 2013), whereas others found that students did not reject more manuscripts with female names (Borsuk et al., 2009; see Lee et al., 2013, for a general review and critique of some gender-bias claims).

Returning to the earlier discussion regarding the advantage of dissecting rather than simply meta-analyzing studies, we note that Tomkins et al. (2017b) conducted a nonsystematic "meta-analysis" of the effects of double-blind review on women's acceptance rates based on their own paper and four others papers that they knew of, all of which are discussed above. Two of these studies (Budden et al., 2008; Roberts & Verhoef, 2016) were shown by later research to have misattributed effects to blind reviewing that were likely due to other factors. Another study in their meta-analysis found an insignificant effect (Blank, 1991), and the fourth study was an experiment of students' reviews (Knoblock-Westerwick et al., 2013). In other words, the highest-ranked effects in Tomkins et al.'s meta-analysis came from studies for which we have raised serious concerns.

Many of the articles on journal acceptances discussed above also tested gender *homophily*—higher evaluations given to one's own gender. Homophily is itself a major topic in the evaluation of grants, journal acceptances, teaching evaluations, recommendation letters, et cetera, that we will defer to a future article. However, we note that if there is widespread homophily, it is surprising that we did not find systematic gender bias in either the narrative analysis or our meta-analyses, because our impression is that the majority of editors and reviewers in many of the scientific fields were men (although we do not have data on this).

We conclude that our meta-analysis and Squazzoni et al.'s (2021) study found only small, statistically insignificant gender differences in the journal acceptance process. The majority of gender differences in acceptance or its components—being sent for outside review (as opposed to desk rejected), reviewer scores, editorial choices after review, et cetera—are not significant at the 5% or even 15% level. When a female disadvantage *was* identified, it was usually small in magnitude. Also, there were occasional instances of female advantage. This does not mean that there was gender parity in every field, time period, and journal; there were occasional gender asymmetries, which we described. Some of these favored male authors, and some favored female authors. However, overall, our meta-analyses and our dissection of key studies revealed no evidence of systematic bias against female authors, notwithstanding claims to the contrary.

Evaluation Context 5: salary

Perhaps no single issue has generated more discussion than the gender salary gap. It is common to read that women earn only 82 cents for every dollar that men earn:

> Research has shown that, compared to female candidates, equivalent male candidates in STEM fields are rated more highly [and] given higher starting salaries. (Dutt et al., 2016, p. 805)

Are male faculty paid more than women in STEM? The literature on academic STEM salary gender gaps, which the economist among us (S. K.) has spent decades analyzing, is sparse and tends to be based on individual institutions or organizations. Moreover, even when national data are reported, they are limited by lack of controls for potentially confounding factors. To confirm this, we conducted a search of Web of Science for articles published between 2000 and 2020 on gender and salaries in tenure stream (tenure track and tenured) academia. Because salary systems outside the United States are more affected by government policies, we limited ourselves to analyses of U.S. data. The majority of studies on the U.S. gender–wage gap in academia were about medicine (mostly contrasting specialties within medicine), which is not typical of academia in general. This left only six articles on the United States, of which one was limited to a specific university and several used aggregated university-level data that were not adequate to control for field, rank, or experience. Only a single article focused on numerous universities and included variables that might be causative, such as objective measures of productivity. We discuss this single exception below—an analysis by D. Li and Koedel (2017), who collected data from six departments in 40 universities from 2015 to 2016.[21]

We begin this section with what many researchers consider to be the most authoritative source of gendered information on faculty salaries, the American Association of University Professors (AAUP) Faculty Compensation Survey, and by discussing new analyses we have done for the present synthesis that control for variables crucial in assessing any real gender gap. This section is therefore briefer than the others because of its reliance on this evidence, as opposed to aggregating studies that share many of the same limitations. Thus, we did not undertake a meta-analysis of salary studies because our own analysis was conducted to improve on prior studies that were limited for reasons we mention below.

Replication and extension of AAUP Faculty Compensation Survey. We began by replicating the single largest, most influential salary study—that conducted by the AAUP—which is considered the gold standard in discussions of gender gaps in faculty salary. Following our replication of the AAUP results, we show why its interpretation needs to be amended by adding missing variables, and we conduct our own regression analyses after adding these missing variables.

The AAUP conducts its Faculty Compensation Survey annually. A summary of the results from its most recent (2018–2019) survey of more than 950 colleges and universities states, "The data also show a significant gender salary gap, as women in full-time faculty positions were paid, on average, 81.6 percent of the salaries of their male counterparts" (Fowler, 2019, para. 2). This figure has become well-known in the popular science media, as in a *Nature* news story that asserts that female faculty are paid only 82 cents for every dollar men are paid (Shen, 2013). However, if one looks beyond the AAUP summary, one discovers a more qualified claim in its report: "The differences are attributable primarily to an unequal distribution of employment between men and women in terms of institutional type and faculty rank" (AAUP, 2019, p. 3). The AAUP results show that within school category (doctoral, master's, baccalaureate, and 2-year schools) and within rank, there were much smaller gendered salary gaps than the 18.4% headline above, which encompassed differing types of institutions and ranks.

The ratio of female to male salaries for full professors ranged from 89.4% in doctoral institutions, to 97.4% in baccalaureate schools, to 99.3% for assistant professors in 2-year schools. Overall, these ratios did not change between 2008 and 2009 and between 2018 and 2019. Thus, much of the pay gap is explained by gender differences in rank and type of school. Having said this, it is important to acknowledge that gender differences in rank and type of school may be at least partially the result of systemic factors that channel women and men into different types of work environments.

Our own analysis shows that if one further divides the faculty by field, even more of the salary variance can be explained, because women and men are unevenly represented in fields that are remunerated the highest. We have analyzed the gender salary gap in academia as of 2017 using the NSF's SDR, a national database. The NSF itself measured and analyzed the gender gap in salaries of PhDs using the SDR in some years of their *Science and Engineering Indicators* (through 2018), controlling for factors such as experience, field, and region, but did not separately analyze the gap in academia. We began our study of gender salary gaps of academics using the SDR by conducting analyses similar to those of the AAUP, which Kahn plans to archive for readers to examine. We found very similar

results as the AAUP, suggesting that we used comparable data and statistical procedures. Specifically, we found that, on average, the gender pay gap for full-time academics is 17.7%, very close to the 18.4% the AAUP analysts reported. However, after controlling for faculty rank (and tenure status) within each school category, we found that the average difference in salaries fell to 9.0%.

After we added controls for experience (years since PhD) and self-reported weekly hours worked, the salary gap fell to 7.2%.[22] Finally, adding category of PhD institution, temporary resident status, and primary work activity (research, teaching, administration) narrowed the unexplained pay gap to 6.9%. There are interesting differences across fields. In all math-intensive GEMP fields—the fields in which women are most underrepresented—unexplained salary gaps are smaller: on average 4.5% in analyses controlling for other factors. In contrast, these gaps are larger in non-GEMP STEM fields—overall: 7.25%; biological sciences: 8.2%.

Gaps were larger before 2000. In analyses differentiating across academic ranks and fields (but no other controls), Ceci et al. (2014) compared academic salaries in 1995 and 2010. They found progress toward salary equalization made in some fields and ranks between 1995 and 2010, but that in three eighths of the field–rank combinations, gaps widened, including significant widening at all three ranks of economists.

We also looked at how gender salary gaps changed over careers. There were smaller gaps in academic salaries for new STEM PhDs—only 5.1% in the math-intensive GEMP fields; in biology, new female PhDs actually earn more than men. However, the more years of experience, the more women's salaries were disadvantaged. Fifteen years after obtaining their PhDs, men earned 8.1% more than women in biology and 4.5% more than women in the math-intensive GEMP fields.

The one article on this topic in our Web of Science search based on many universities and fields, with data on individual faculty and their productivity, was by D. Li and Koedel (2017), who collected data on six departments in 40 universities between 2015 and 2016. With no controls, the gender pay gap was $23,320, or 19%, close to the AAUP's 18.4%. When analyses controlled for university and field, this dropped to 12%; when they controlled for experience and PhD school, this dropped to 6.1%; and when they further controlled for research productivity, the gender pay gap dropped to $4,280, or 3.6%. The results without the productivity controls are close to what we found in the SDR. Therefore, we believe that the 3.6% figure represents a good approximation of the pay gap that would remain after one controls for research productivity—a gap 80% smaller than the average gap without controls of 18% to 19%.

However, even small gaps can translate into nonnegligible lifetime earning gaps. For example, in 2014, female assistant professors in psychology earned, in inflation-controlled dollars, approximately 4% less than male assistant professors ($68,640 vs. $65,900). This became an approximately 7.2% gap among full professors ($110,145 vs. $102,165; APA Committee on Women in Psychology, 2017). As Gruber et al. (2021) note, the gap is even larger if we include employer contributions to retirement savings. Thus, a small initial gender pay gap that cannot be accounted for by differences in factors such as productivity, type of institution, hours worked, and field represents the seeds of lifetime inequality.

What causes these gender gaps in salary? Some of the unexplained gender salary gap may be due to implicit bias (although this seems unlikely in biology, where starting salaries are higher for women), and some of it may be due to differences in willingness to negotiate and solicit outside offers, as we address below. However, as D. Li and Koedel (2017) showed for their sample, given the substantial and pervasive gender gaps in publications that we review below in Context 7, a substantial percentage (there, 41%) of the gender pay gap is due to productivity differences (for earlier support, see Perna, 2001). Relatedly, some of the remaining pay gap may be due to women's work discontinuities for family leave (e.g., Huang et al., 2020; Morgan et al., 2021) or to a desire to keep jobs flexible (Goldin, 2014). (However, in the SDR national data, women with children have higher salaries, all else being equal, which is probably not causal but indicates positive selection—the most able female scientists get married and have children—similar to what we found for married male scientists with children.) Finally, some of the relatively small remaining pay gap may be due to women's lower likelihood of negotiating higher salaries or their lower likelihood of pursuing more lucrative job offers. The lower likelihood of negotiating higher salaries may itself be due to bias.[23] Without specific data on family leaves, past employment, and job pursuit, it is impossible to know how much, if any, of the less than 4% unexplained pay gap is attributable to bias.

In sum, we conclude that the evidence supports the claim that women are paid less than men in tenure-track academia, although the magnitude of the gap is much smaller (60%–80% smaller) than often claimed in executive summaries and headlines, and in some situations has disappeared. We were able to closely replicate the AAUP findings using a different representative data set and the controls that they had and then add additional moderators (PhD institution, hours, experience) that had not been factored into their analysis (nor in most other previous analyses). This new evidence, along with D. Li and Koedel (2017) and Ding et al.

(2021) further give us a sense of how much of the gender pay gap may be due to productivity differences. In conclusion, although we identified salary as one of two domains (the other being teaching) in which gender bias was found, the magnitude is much smaller than the commonly claimed 18% gender pay gap (Shen, 2013). None of this means that there were not larger gender gaps in salaries during earlier periods.

Evaluation Context 6: recommendation letters

One factor frequently invoked to explain why women are underrepresented in GEMP is gender differences in recommendation letters:

> Women . . . receive less compelling letters of recommendation. (Witteman et al., 2019, p. 531)

> In letters of recommendation for faculty positions, studies have found that women are more likely to be praised for their "communal" skills (i.e., collaboration), whereas men received more mentions of "agency": being "brilliant" at research or a "genius." (Weisshaar, 2017, p. 535)

> Female candidates are half as likely as male candidates to receive an excellent letter or to have 'standout' adjectives like 'excellent,' 'outstanding' or 'extraordinary.' (Grogan, 2019, p. 4)

In a survey of nearly 100 chairpersons, letters of recommendation were regarded as highly important in evaluating applicants for faculty positions (Sheehan et al., 1998). There is a considerable literature describing how, in general, letters of recommendation for men and women differ both in their content and, at times, in the differential interpretations that various readers make of the same letter. These studies often involve asking students to interpret letters crafted to invoke gender or racial stereotypes (e.g., Biernat, 2012; Biernat & Eidelman, 2007).

In contrast to this larger literature on experiments using student raters or letters written for nonprofessorial positions, the literature examining letters of recommendation for professorial jobs is much smaller. Here, we examine this literature, first describing the myriad linguistic codes that have been used to study gender bias in letters of recommendation.

In the general (i.e., nonprofessorial) literature on letters of recommendation, studies sometimes contrast the frequency of agentic and communal words, constructs borrowed from social-role-congruity theory (Eagly et al., 2000; Eagly & Karau, 2002). *Agentic* terms imply active, take-charge leadership characteristics (e.g., "self-confident," "logical," "goal oriented," "assertive," "ambitious," "independent"), whereas *communal* words connote other-oriented concern (e.g., "nurturant," "sensitive," "cooperative," "relationship oriented," "sympathetic"). A body of research (e.g., Biernat, 2012; Madera et al., 2009) demonstrates widespread cultural stereotypes depicting women as communal and men as agentic and implies that women are less competent than men in male gender-typed domains that associate competence with agentic traits). Agentic and communal words have been examined in three tenure-track studies (Bernstein et al., 2022; French et al., 2019; Madera et al., 2009)[24] with differing results.

Other researchers have contrasted gender differences in the use of *standout* adjectives that convey exceptional talent (words such as "outstanding," "amazing," and "unrivaled") and *grindstone* words (such as "hard-working," "diligent," and "reliable"). Blue and her colleagues (2018) argue that grindstone words carry the implication that for women, effort compensates for their deficiencies in ability:

> "Standout" words, which portray a candidate as talented and exciting, are most often found in letters of recommendation for men. Grindstone words, which create the impression that a candidate works hard but is not intellectually exceptional, are more often used for women. (p. 42)

Still other studies have examined gender differences in positive and negative emotion words ("pleasant," "critical"), whereas others have examined power and achievement words ("accomplished," "commanding"). Finally, some studies have examined *doubt-raising* words that damn with faint praise, such as, "She is very solid but unlikely to become a superstar."

Are these gendered differences in word associations that have been observed in experiments and surveys also found in letters of recommendation for academic tenure-track applicants? Are letters written on behalf of male applicants for tenure-track positions more likely to depict them as brilliant, logical, and confident and letters written for female applicants to depict them as cooperative, sympathetic, and diligent?

To evaluate these questions, we conducted a comprehensive search on both Web of Science and Google Scholar for studies on applicants for academic tenure-track (or equivalent) jobs between 2000 and 2020. There have been only nine such studies, confirmed by checking all references to each of these nine studies for additional studies. These nine studies are listed with their basic findings in Table 4. As the table shows, the sample sizes of studies on letters for tenure-track appli-

cants ranged from small (*N*s ≤ 300 letters) to large (*N* > 2,000 letters), and there are only a few dimensions that these nine studies have in common; notably, word length, standout words, and doubt-raising words. Thus, different teams of researchers (Table 4) have contrasted gender differences in the use of standout adjectives that convey exceptional talent and grindstone words such as hard working and persistent. Three studies have examined *ability* words (Bernstein et al., 2022; S. Li et al., 2017; Schmader et al., 2007), which some scholars claim distinguish women and men, with the latter being alleged to possess more native raw talent ("brilliant," "genius," and "gifted"; see Leslie et al., 2015, for survey of faculty beliefs about the role of brilliance in male-dominated fields). Two of the studies in Table 4 used word length as a covariate, whereas six of the remaining seven found that word length was comparable in letters written on behalf of men and women.

The data in these studies were collected during the mid-90s to 2017, and they differed in the use of controls (e.g., Trix & Psenka, 2003, had access only to letters written on behalf of successful applicants and none written on behalf of the larger group of unsuccessful applicants, so there is no information about whether the letters written for unsuccessful applicants were biased). The few studies with similar dependent variables were from different fields, and no two studies were comparable in terms of all relevant factors. Thus, a meta-analysis of these nine studies is contraindicated because of the heterogeneity of dependent variables, fields (medicine, physics, biology, psychology, sociology geology), epochs (mid-1990s to 2017), formats (standardized vs. free-style narrative), controls (length, status of writer), and moderators. Fortunately, despite the differences across the studies, when we calculated effect sizes for the different dependent variables, they pointed in the same direction, with some exceptions: The findings reveal no systematic gender bias. For example, the only study to have found greater letter length in letters written for men was one of the smallest (Trix & Psenka, 2003; *N* = 300), and other studies found the opposite or, most often, no difference.

Dutt et al. (2016)[25] found that male candidates for postdoctoral positions in geoscience at Columbia University were twice as likely as women to receive excellent letters, in the sense of having more standout words. However, they too found that the genders were the same in some other dimensions. Moreover, their study lacked strong controls for characteristics of applicants. If members of one gender have superior accomplishments (e.g., greater productivity), this might influence writers' word choice. The few studies that have attempted to control for applicants' characteristics have done so by covarying the candidates' numbers of publications, institutional prestige of their PhD, and numbers of presentations (Madera et al., 2009; Schmader et al., 2007). Even these controls might miss other information, such as writers' knowledge of the candidate's contribution to each publication, status of letter writer (rank), and impact of journals in which applicants publish.

Overall, Table 4 shows that there is no compelling evidence for the assertion that letters for women (compared with letters for men) are shorter, more communal, and less agentic; contain more doubt-raising words or more negative emotions; or are written by lower-status authors. For every study that has documented a gender bias in one of these categories of words, there is at least one that has found equivalent or even larger effect sizes for no gender bias, or even reverse bias. For example, although Madera et al. (2009) found that letters written on behalf of female applicants contained more communal (nurturant, kind, sensitive) and less agentic (self-confident, independent) words (and that communal words are negatively correlated with hiring decisions), two other studies did not, including the largest one.

Support for this conclusion is underscored by the largest study. Bernstein et al. (2022) found only a few gender differences, most favoring female applicants, in an analysis of 2,206 recommendation letters in physics and psychology. When employing the same type of analysis that previous studies used, *Linguistic Inquiry and Word Count* (LIWC; Pennebaker et al., 2015), these researchers found that neither the gender of the applicant, the gender or rank of the writer, nor the discipline mattered for six of seven dependent variables (i.e., letter length, agentic words, communal words, standout words, grindstone words, achievement/power words, positive emotion words). Using the traditional LIWC analysis, Bernstein et al. found that the few significant differences were as likely to advantage female candidates as their male counterparts (e.g., there were more positive emotion words and fewer negative emotion words in letters for women; in addition, there were generally longer letters for women, and length is characteristic of stronger letters). Using a novel bottom-up analysis that focused on individual words rather than analyses that validated top-down word lists, such as LIWC, which have been used by previous researchers, Bernstein et al. found more mentions of the words "physicist," "intellect," and "creative" in letters for men, but they found three times more mentions of the key word "brilliant" in letters for women. Overall, there were not fewer standout or agentic terms in letters for women, nor were there more communal terms for women, as some scholars have opined (and as suggested by the quotes opening this section).

On the basis of our analysis of the nine studies in this domain, we conclude that no persuasive evidence

Table 4. Studies of Academic Letters of Recommendation

Variable	Li et al. (2017)	Trix & Psenka (2003)	French et al. (2019)	Madera et al. (2019)	Madera et al. (2009)	Messner & Shimahara (2008)	Schmader et al. (2007)	Dutt et al. (2016)	Bernstein et al. (2022)
Data years	2015–2016	Mid-1990s	2015–2016	1998–2006	1998–2006	1998–2006	Not specified	2007–2012	2010–2017
Field	Medicine	Medicine	Medicine	Psychology	Psychology	Medicine	Chemistry, biochemistry	Geoscience	Physics, social science
Sample size	237	300	559	624	624	763	886	1,224	2,206
Controls?[a]	Prescreened	No	Prescreened	Yes	Yes		Yes	No	Yes
Word length	Genders same	More male long letters	Genders same	Controlled; n.s.	Controlled; n.s.	Genders same	Genders same	Genders same	Genders same
Word type									
Standout	Genders same	More multiple standout words for men	More for women			Genders same	More for men	Double for men	Genders same
Agentic			Genders same		More for men				Genders same
Communal			Genders same		More for women				Genders same
Grindstone	Genders same	More for women	Genders same				Genders same		More for women[a]
Doubt-raising		More for women	Genders same	More for women			Genders same	Genders same	Genders same
Positive, negative emotion	Genders same								More positive for women, more negative for men[a]
Ability	More for women						Genders same		Genders same

Note: [a]These findings interacted with either the field or gender of the writer.

exists for the claim of antifemale bias in academic letters of recommendation. The empirical findings in Table 4 support a mixed pattern with a few gender differences but mostly gender neutrality, and this is underscored by the largest study, which is nearly twice as large as the next-closest-sized study (Bernstein et al., 2022). Finally, although there may appear to be temporal trends across these studies, with earlier ones showing the most bias and latter ones the least, perhaps because of writers becoming more skilled in reducing (or masking) bias, there are no reliable time trends in the past two decades, although there is some suggestion of earlier bias: The oldest study's authors, Trix and Psenka (2003), gathered their data between 1992 and 1995, and they reported the largest gender bias. However, they also had one of the smallest samples and no control group of letters for unsuccessful applicants, which limits their estimate of bias because letters written on behalf of failed applicants may be unlike those written for successful ones. All other studies were conducted later and found either mixed gender differences or no gender differences.

Also, there does not appear to be a linear time trend over the past two decades because the two studies by Schmader et al. (2007) and Madera et al. (2019) found some bias and some lack of bias, whereas in the same 2007 to 2009 period, Messner and Shimamura (2008) found no bias for grindstone words, standout words, et cetera. Dutt et al.'s (2016) study 7 to 9 years later found substantial bias against female applicants for postdoc positions at Columbia University, but Bernstein et al.'s (2022) study, which collected data over the 2011 to 2017 period that overlapped most of Dutt et al.'s data, found no temporal trend in bias. In sum, there may be some suggestion of gender bias in letters prior to 2000 (Trix & Psenka, 2003) but no systematic evidence after 2000.

Context 7: gender gaps in research productivity

Research productivity is not an evaluation context per se. However, because it possibly mediates academic labor-market evaluations—for hiring, grants, salary, tenure and promotion, and letters of recommendation—we provide a review of reviews here. We summarize the research[26] on gender differences in publishing productivity, emphasizing the largest and most comprehensive analyses and reviews. As has been noted throughout, very few studies of gender gaps in salary, grants, hiring, and letters have controlled for productivity, but the case for doing so seems obvious, given that what professors actually produce by way of tangible written work is arguably the key currency of their profession. At the outset of this section, we note that controlling for quantity is not the same as controlling for quality—and direct tests of quality are rare. Notwithstanding this caveat, the dominant view is that women publish less than men:

> One of the most consistent findings in the literature on research productivity is that women tend to have somewhat lower publication rates than men. (Abramo et al., 2007, p. 518)

> A consistent finding in the literature is that male researchers generally do indeed publish more than women. This pattern has been revealed across numerous countries, fields, and time periods (Cole & Zuckerman, 1984; Fox, 2005; Long, 1992; Xie & Shauman, 1998, 2003; Abramo [et al.], 2009; van Arensbergen [et al.], 2012; Larivière [et al.], 2013; Elsevier, 2017). (Abramo et al., 2021, p. 102)

Key finding of lifetime gender gap in productivity. Beginning with Cole and Zuckerman's (1984) seminal analysis of gender differences in productivity, many studies have documented that over their lifetime, academic women publish less than academic men. Studies fall into three categories: (a) surveys based on self-reported publications; (b) administrative records from individual universities, granting agencies, or journals; and (c) bibliometric analyses of the published corpus of research over some specified time period (e.g., 2 years, 5 years, 10 years, lifetime), the latter enabled by computerization of records in databases such as Web of Science, Scopus, Medline, dblp (an online computer-science bibliography), and JSTOR.

As will be seen, the vast majority of these studies (including all of the massive comprehensive bibliometric analyses) have found substantial productivity gaps revealing that men publish more articles than women. However, there are a number of factors that need to be considered that complicate this simple conclusion. For example, the different studies are based on very different populations (ranks of authors, disciplines, nations, career points), very different dependent measures (e.g., annualized publications, lifetime publications, publications in high-impact journals, first-authored/any-authored, most cited), different epochs, different cohorts, and noncomparable analytic methods. To make comparisons even more problematic, studies combine different but overlapping subsets of fields. Such heterogeneity in dependent variables and analytic methods obscures interrelationships among variables and can result in erroneous inferences and spurious conclusions (Lipsey, 2003; Stone & Rosopa, 2017). However, these studies point overwhelmingly in the same direction: greater male lifetime productivity.

One major dissimilarity is that bibliometric and journal-based analyses include only scientists who have at least one publication within the window of time analyzed, thus excluding scientists who have published no articles during that period. We have analyzed the publications of STEM PhDs in the NSF SDR—only available through 2008—and found that up until that time, women were more likely than men to have zero publications in a given 5-year period, a finding that others have also reported (e.g., Fox, 2005, reported that in her STEM sample, 8.7% of women vs. 7.1% of men had zero publications and 10.1% of women vs. 3.3% of men had one publication). Consequently, bibliometric analysis depends on a very different sample than survey data and yields more gender-equal publication records, because it includes only those scientists who are research active in a specific epoch. Another distinction is that cumulative publication records covering long periods will find more gender inequality, because women experience more career interruptions and have shorter overall research careers. Finally, surveys of faculty will miss former employees who are not currently employed. Any attempt at aggregation must take into account these and many other differences among published analyses.

Changes over time in gender publication gap. Several U.S. surveys reported time trends in gender publication gaps. Xie and Shauman (1998) compared faculty publications over consecutive 2-year windows in nationally representative cross-sectional surveys to identify time trends from 1969 to 1973, 1973 to 1988, and 1988 to 1993. Without controls, the gender publication gap started at 37% in 1969 and fell to about 26% by 1993. With controls for experience, rank, university type, and other factors, the gap fell to 9.6%. Perna (2001) used the same 1993 NSF survey as Xie and Shauman did for 1993, except that she limited analysis to refereed articles and analyzed by rank. She found that the gender productivity difference was 16% for assistant professors, only 3% among associate professors, and 12% for full professors with 13 to 20 years of experience.

Ceci et al. (2014) used a different national survey, the SDR, to study refereed articles in two 5-year periods: 1991 to 1995 and 2004 to 2008. They found that from 1991 to 1995, the male publication advantage without controls was 22.5%, similar to Xie and Shauman's (1998) estimate of 26% without controls. By 2004 to 2008, the average gap decreased to 19.6%. Ceci et al. also found that the level and trend in the productivity gender gap differed by seniority, falling from 24.5% between 1990 and 1995 to 6.2% between 2004 and 2008 for assistant professors, from 24.1% to 20.0% for associate professors, and from an 11.6% gap in 1990 to 1995 to insignificantly higher women's average publications in 2004 to 2008 for full professors.

The major bibliometric study that allows us to evaluate time trends is by Huang et al. (2020). They reconstructed the complete publication history from Web of Science of more than 1.5 million authors from 1955 to 2010. This study defined its population differently than other studies: It included all authors with at least two publications between 1955 and 2010 and whose publishing careers ended between 1955 and 2010. Thus, they not only excluded people with zero and one publications but also excluded everyone who published anything after 2010. This means their analysis excluded those scientists who are currently research active.

Huang et al. (2020) found considerable gender publication gaps over authors' research lifetimes (from their first to last publication), with a male advantage of 27% (13.2 papers vs. 9.6 papers). If one looks at the time trends from scientists who ended their research careers in the 1950s to those who ended them in 2000 to 2010, one sees several patterns. First, for both men and women, total productivity over their entire publishing careers increased considerably. The average went from around 5.6 (1950–1959) to 13.9 (2000–2009). Over this time, the gender gap increased much more than proportionately over the decades, so that "the gender gap in total productivity rose from near 10% in the 1950s to a strong bias toward male productivity (35% gap) in the 2000s" (Huang et al., 2020, p. 4613). Yet there were negligible annual productivity differences between men and women (women: 1.33 per year, men: 1.32 per year) that showed no time trend. These two trends seem contradictory, but are due to women having considerably shorter publishing careers than men. Huang et al. found that men on average published over a span of 11.0 years, whereas women published over a span of 9.3 years (a 16% difference), so that "women scientists have a 19.5% higher risk [of leaving] academia than male scientists, giving male authors a major cumulative advantage over time" (p. 4613). Over the decades, careers were getting longer, but the difference between the genders grew substantially, so that women's careers were 7.9% shorter than men's in the 1950s but 21.9% shorter in the 2000s.

Another major bibliometric study measuring time trends in gender differences in publications is Elsevier (2017), which mined their Scopus database of 62 million documents from 1996 to 2000 and 2011 to 2015. These researchers measured the number of publications for each person who had at least one publication in each 5-year period for 12 "geographies," comparing 1996 to 2000 with 2011 to 2015. As a bibliometric study, it excluded researchers with zero publications in these 5-year spans, and thus found smaller gender gaps than

studies that did not exclude researchers who did not publish, because women are more frequently in the zero-publication category. Indeed, the gap that Elsevier identified for the United States in 1996 to 2000 was considerably smaller than the gaps estimated by Ceci et al. (2014) for either 1991 to 1995 or 2004 to 2008. Overall, Elsevier found that in most countries, women published fewer papers on average between 2011 and 2015 than they had between 1996 and 2000, whereas men published more. For the United States, it found that the gender difference in publications doubled from 5% between 1996 and 2000 to 10% between 2011 and 2015.[27]

Field differences. As we have found with many other aspects of scientific careers, there is generally no one-size-fits-all result across fields, with heterogeneous gender publication gaps. Thus, we conclude by highlighting some notable field differences. Biology—where women now make up more than half of new U.S. PhDs and the field with the most women in Huang et al.'s (2020) multidecade census of authors—is also the field with the largest female disadvantage in annualized publications (7.5 ppt) and in total productivity (37.7%). It also shows the second largest difference in career lengths (19.6%). Ceci et al. (2014) found that between 1990 and 1995, biology/life sciences had one of the largest gender gaps among assistant professors and full professors (each 35%). However, by 2003 to 2008, this gap fell (to < 20%), although it was still significant. Symonds et al. (2006) estimated publication differences among 168 life scientists (United Kingdom and Australia) during 1993 to 2005 and found a similar gender difference in long-term productivity of 40%. The gender productivity gap begins in graduate school (Schaller, 2023), as we describe below.

Psychology also has large gender publication gaps (a 4.6-ppt male advantage in annualized publications in Huang et al., 2020, and a 23.5-ppt advantage in total productivity). Moreover, Ceci et al. (2014) found that psychology's gender imbalance among assistant professors grew from 25% in the years 1990 to 1995 to 34% in 2003 to 2008. Abroad, van den Besselaar and Sandström (2016) found that Dutch midcareer men outperformed women in publications by 63% in psychology. Odic and Wojcik (2020) analyzed authors in the top 125 journals of psychology from 2003 to 2018 and found that controlling for seniority, published men had a 22% publication advantage over published women (see also Aguinis et al., 2018; D'Amico et al., 2011; Madison & Fahlman, 2021).

Madison and Fahlman (2021) analyzed a representative sample of faculty reaching the professor level at the largest Swedish universities in six disciplines (psychology, linguistics, political science, social sciences, law, and medicine). They evaluated whether men or women had more publications and citations at the point that they were appointed to a professorship. Overall, men had more publications and citations, although when broken down by field, greater male productivity reached significance in only three fields (social sciences, linguistics, and education). (Interestingly, Madison and Fahlman interpret the lower cumulative productivity of women at the time they got tenure as pro-female bias.) Bird (2011), on the other hand, did not find a gender productivity gap in psychology or social policy fields in the United Kingdom, whereas she did find gaps in other UK social science fields.

In contrast to the above LPS fields, in the math-intensive GEMP fields, although women comprise a smaller percentage of scientists, the gender differences in publications appears to be smaller and shrinking. Huang et al.'s (2020) bibliometric analysis of annualized publications in these fields showed either female advantages or very small male ones (2.6 ppt, 5.2 ppt, and 2.1 ppt female advantage in engineering, computer science, and physics, respectively, and 0.8 ppt male advantage in math), with career lengths similar to the overall average. In Ceci et al. (2014), the publication gaps in the GEMP fields engineering and math/computer science started out substantial and significant in 1990 to 1995 for assistant professors but became insignificant by 2003 to 2008. Among associate professors and full professors, the gaps in these fields were small and insignificant in 1990 to 1995, and by 2003 to 2008 there were female advantages in math/computer science (only significant for associate professors). And among associates in engineering, Duch et al.'s (2012) study of 4,292 faculty in top U.S. research universities also found a similar ordering of fields: Gender publication gaps in engineering were smaller than in psychology, which in turn had smaller gaps than in biology. Thus, the patterns across studies are similar.

In social sciences besides psychology, two fields stand out as opposites. Political science had a large female advantage in annualized publications and a 3.3-ppt female advantage in total impact (Huang et al., 2020). (Neither Ceci et al., 2014, nor van den Bessalaar and Sandström, 2016, separated this field from the other social sciences.) In contrast, Huang et al. found that in economics, men had a 28% productivity advantage in early careers and a 50% advantage in midcareers. Ceci et al. found that the productivity gap among economists increased from 1990 to 1995 to 2005 to 2008 (from 22% to 52%), a concerning trend. (Huang et al. did not separate out economics.)

In sum, gender productivity differences are smallest in GEMP fields (with the exception of economics) and are largest (and possibly growing) in biology, psychology, and economics.

Publication gender gaps over careers. A number of surveys show that the gender publication gap starts in graduate school. Lubienski et al. (2018) examined the publications of approximately 1,300 graduate students at a large R1 university and found that men submitted 59% more articles than women and published 69% more (by our calculation, these represent effect sizes of $d = 0.54$ and $d = 0.53$ greater male productivity, respectively), with differences being larger in natural sciences and engineering than in humanities, education, or social sciences. Pezzoni et al. (2016) reported on publications of 933 U.S. graduate students (PhD cohorts 2004–2009) at California Institute of Technology and found that female students published 8.5% fewer articles, with the greatest difference in biology (13%) and the smallest in physics (5.5%). Feldon et al. (2017) analyzed data from 100 U.S. graduate programs in biology and reported that male first-year PhD students published 15% more articles than females (2014–2016, self-report), despite women reporting that they worked more hours. In the largest study of gender productivity gaps among graduate students, Schaller (2023) analyzed productivity data from 42,922 doctoral students at 235 institutions. Men published 10% more first-authored papers and 15% more total papers, and this male advantage appeared early in their graduate careers and was not moderated by the gender of student-advisor dyads or advisor productivity. In the United States in social sciences, Lubienski et al. (2018) found that average publications of male and female graduate students were equal in psychology, but in economics, publications of men were 75% higher.

The publication gap continues during postdoctoral research. In a large-scale survey, Davis (2009) found that male STEM postdocs published 34% more than female STEM postdocs. However, van Arensbergen et al. (2012) found that among Dutch scientists in the 3 years after obtaining their PhDs, there was no overall significant gender difference in publications (also confirmed by van den Besselaar & Sandström, 2016).

In early careers, women's productivity is likely often related to childbearing and child-rearing (Williams & Ceci, 2012). Several analyses have not found pronounced motherhood gaps, such as Sax et al.'s (2002) analysis of 8,544 full-time faculty who were part of a national faculty self-report (see their Table 3 for lack of motherhood effect). However, using bibliometric databases of publications, Morgan et al. (2021) surveyed a representative sample of GEMP faculty (computer scientists) as well as faculty in two non-GEMP fields (business and history), stratified by faculty rank and departmental prestige. They linked faculty self-reported productivity to publication data in dblp, an online computer-science database, to analyze 79,274 publications of 1,061 computer-science respondents. A substantial gender publication gap was evident: In the decade following the birth of their child, female computer science faculty "produce on average 17.6 fewer papers than fathers—a gap that would take roughly 5 years of work for mothers to close" (p. 5).

However, they also noted that this motherhood penalty had relatively little impact on later productivity, when women returned to their premotherhood levels of authorship. This finding is echoed in Huang et al.'s (2020) massive analysis and also by Cameron et al.'s (2016) analysis of productivity within the field of ecology, where women had career interruptions but comparable productivity with men during their research-active periods.

Morgan et al. (2021) also contrasted gender publication gaps of faculty without and with children, as summarized in Figure 6, reproduced from their article. As the figure shows, among computer-science faculty without children, women have somewhat lower publication productivity than men (male faculty cumulatively published 5.2 more papers than female faculty by their 10th year after being hired as assistant professors), but women pay a large motherhood penalty for the decade surrounding a child's birth (with a publication gender gap of 11.9 papers during these 10 years), resulting in a substantial lifetime productivity gap. This translates to female computer-science faculty with no children publishing 87.6% of the total number of papers that their male counterparts published, but computer-science faculty who are mothers publishing only 73.6% as many papers as fathers over the same period.

As Morgan at al. (2021) noted, "fathers experience less of a parenthood penalty than mothers, because men may have nonacademic partners, leading to more flexibility in adapting to an abrupt change in time available for research due to caregiving" (p. 5). This cultural context of structural gender roles is relevant in interpreting productivity gaps and should not be interpreted to mean that men, who have higher lifetime productivity, are entitled to greater academic rewards. We must remain cognizant of and actively study the caretaking roles occupied primarily by women (e.g., caring for children and aging parents) and the consequences for productivity gaps. The causal forces of gender gaps and efforts to mitigate them may be far more complicated than this. This point was underscored by Eagly (2020).

Country differences. Several studies, mostly bibliometric ones, broke down productivity by country. Elsevier (2017) found that in all countries except Japan, the gender gap started small in 1996 to 2000 but increased by 2011 to 2015. For instance, in the 28 European Union countries as a whole, the gap rose from 4% to 13%; in the United Kingdom, it rose from 8% to 25%. In Canada, it

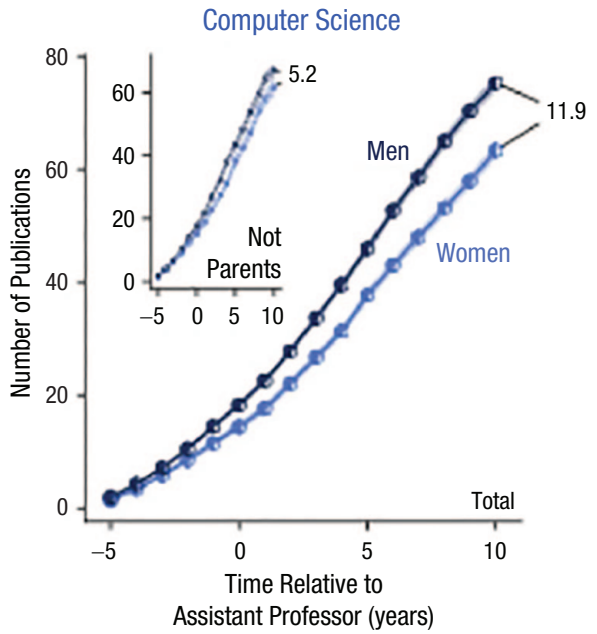

Fig. 6. Productivity of computer-science faculty, measured by mean number of publications as a function of the time of faculty's first tenure-track position (reproduced from Morgan et al., 2021, Fig. 2a). The main graph shows annual productivity of faculty members with children. The inset graph shows a counterfactual case of a faculty member without children who is aligned on key variables such as chronological age and career age. In each graph, results are shown separately for male and female faculty.

also rose from 9% to 24%, and in Brazil, it rose from 6% to 20%. In Japan, the gap remained stable around 30%. (This is quite different from the 56% Japanese productivity gap measured by Aiston & Jung, 2015, based on self-reported STEM academic productivity derived from the 2008 Changing Academic Profession survey, which included ~1,500 academics.) Mayer and Rathmann (2018) found a significant gender gap in publishing journal articles (but not chapters or books) among German full professors, which also was found by Ceci et al. (2014) for American full professors.

Huang et al. (2020) and their appendix indicates substantial heterogeneity across countries: The United States had a 6.6% male advantage in lifetime publications; this number was 7.3% in Canada and 9.5% in the United Kingdom, and other Western European countries had similar gaps. Some Eastern European and African countries had female advantages (e.g., Serbia: 25.0%, Kenya: 22.7%). Another large-scale bibliometric study was reported by Larivière and his colleagues (2013) of nearly 5.5 million research papers in the Web of Science between 2008 and 2012. Women accounted for a smaller proportion of fractionalized authorships at all ordinal positions (first author, last author, sole author). For example, for every first-authored publication by a woman, there were 1.93 first-authored articles by a man. Greater male productivity was observed across the most developed countries; countries with female dominance tended to be former communist countries and countries with overall low productivity (Macedonia, Sri Lanka, Latvia, Ukraine, and Bosnia and Herzegovina).

Most of the gender publication gap occurs in the top tail. Despite the large differences in the mean number of publications per person in these large-scale analyses, Huang et al. (2020) found that there was no gender difference in median publications per person. The difference was entirely due to the top tail, where men dominate among authors with very large numbers of publications. Other studies also found men more likely to be top publishers: Kelchtermans and Veugelers (2013) found that men were approximately 3 times as likely as women to be persistently top performers in science at a single university, similar to what Odic and Wojcik (2020) and Aguinis et al. (2018) found for psychology. And Braisher et al. (2005) found that the male advantage that exists among authors who published in top journals from 1999 to 2005 resulted from a small number of very successful men who had more than 30 publications between 1999 and 2004. Although no woman had more than three publications in *Science* and *Nature* during this period, 13 men had four or more publications.

Brower and James (2020) also found that men in New Zealand were overrepresented at the rightmost tail of productivity. And Sax and her colleagues (2002) found in their analysis of a national faculty survey that although the gender gap in productivity had been cut in half by the turn of this century (and completely erased among the least productive), men still were disproportionately among the most productive, with a higher percentage of men having three or four publications over a 2-year period (18.8% vs. 15.6% for women), and men had nearly double women's rate of five or more publications (16.6% vs. 8.8%, respectively). Aguinis et al. (2018) also found that the productivity gap was most pronounced among the "elite" faculty (the top 10%, 5%, and 1%). Abramo et al. (2021) found that among Italian and Norwegian faculty, most of the gender productivity gap was due to the top 10%, with very small gender gaps among the other 90%:

> In line with a few previous studies, our results reveal that most of the overall gender differences in both countries can be explained by the tails of the distributions—particularly a much higher proportion of men among the top performing scientists. Findings confirm previous results by Abramo, D'Angelo, Caprasecca, et al. (2009), who found . . . the performances of men and women . . . do

> not differ much, except in the top performing groups. (p. 14, citations omitted)

In sum, men are almost always overrepresented among the most productive, with a narrowing of the gender gap as one moves away from the right tail toward the mean.

Summarizing findings on gender gaps in productivity. The bottom line is that although there are substantial differences across countries, epochs, ranks, and fields, the evidence from more-developed countries shows that men on average publish more than women in science over their lifetimes and tend to have higher levels of publications and other productivity measures (Astegiano et al., 2019). This is partially the result of men remaining active and publishing and working more continuously (with fewer interruptions for family reasons); and relatedly, it is partially due to some women not publishing at all in given 5-year periods, and it is exacerbated by a small percentage of very productive researchers who are disproportionately male. The effect of having children in the first decade after being hired accounts for much of the gender variance, although even among nonparents, men outpublish women, and over their lifetimes, the gap is substantial (Morgan et al., 2021; cf. Sax et al., 2002). Although the gender gap seems to have narrowed from the 1960s, it may have risen in the past two decades in some fields at some career stages. This difference can have large impacts.

To end this section at the beginning, we note that persistent gender differences in productivity could be contributors to potential gender differences in hiring, grants, letters, and especially salary in the tenure-track academy. However, none of this implies that women are doing less work overall than men. It has been documented that women assume greater responsibilities and not just at home. Some authors have argued that women teach and do more service than men. However, the evidence for such claims is contradictory and cannot explain the hiring data (or possibly any other data). For example, the NRC (2010) data, which is based on all hiring at 89 research universities in the mid-1990s, found no gender differences in teaching or service loads, and Davis's (2009) national survey of postdocs also showed no gender differences in teaching and service. Other authors have documented that women are not acknowledged for their work on joint projects as often as men are (Ross et al., 2022), although this seems unlikely to be the cause of the overall productivity gaps described above, given that women are also sole authors less often than men and during their active work life, women publish as often as men, which would not be the case if their participation was not being acknowledged. Although we have not comprehensively studied this issue, it seems likely that the cause of productivity differences entails other considerations, most likely women's greater family responsibilities, which have been amply documented (e.g., Xu, 2015).

Discussion

As seen in the high-profile quotes throughout this article, claims of gender bias in the academy have been omnipresent, appearing in the most prestigious journals. High-level reports by the National Academies of Sciences, Engineering, and Medicine (NASEM, 2018, 2020a); articles in *Science*; and editorials in *Nature* claim that gender bias exists in all aspects of STEM academia—without qualifications. For example, the National Academy of Science's consensus report recently stated: "Bias, discrimination, and harassment are major drivers of the underrepresentation of women in science, engineering, and medicine" (NASEM, 2020b, p. 1). Writing in *Science*, Fortunato et al. (2018) asserted that "women have fewer publications and collaborators and less funding, and they are penalized in hiring decisions when compared with equally qualified men" (p. 3, citations omitted). And in an editorial in *Nature*, Urry (2015) claimed that "every major criterion on which scientists are evaluated . . . has been shown to be biased in favour of (white) men" (p. 472).

In response to such influential, omnipresent assertions, we set out to synthesize all the empirical evidence in six key domains that are important for women early in their tenure-track careers and that have been the focus of claims of bias. We evaluated what each study's data actually revealed—which we sometimes discovered was different from what its authors or its abstract asserted. Our conclusion is that in four of these domains, claims of widespread gender bias are not supported. Rather, these claims rest on selectively chosen evidence and ignore important counterevidence and sampling and methodological limitations. In two domains, we concluded that claims of bias are supported.

Before we summarize our findings below, we reiterate a caveat noted throughout this article: The failure to support specific claims of bias does not deny the possibility that broader, systemic barriers against women in the academy exist and/or that significant bias existed before 2000. We did not examine systemic claims of bias, such as the tenure schedule that imposes inflexible time-career paths or structural societal norms that burden women with greater responsibilities outside of their academic jobs or that penalize women for negotiating forcefully for wage increases or seeking outside offers. Other scholars have identified a myriad of such

systemic barriers. But when it comes to specific claims about biased grant reviewers, search committee members, journal editors, and letter writers, the claims of antifemale bias were not supported, and in one case (tenure-track hiring), the data actually supported the opposite conclusion—that of pro-female hiring bias. This pro-female hiring advantage has continued after the closing of our inclusionary period, 2020 (Henningsen et al, 2021; Solga et al., 2023).

Evidence of gender bias in two domains

We found that there is evidence of gender bias in the domains of teaching ratings and faculty salary. In the former, on the basis of our dissection of key studies in this domain, we concluded that there is persistent gender bias in teaching ratings. After scrutinizing key studies, we found that women professors were more likely to be unfairly evaluated than their male counterparts, although the magnitude of this bias is unclear, except in the case of word frequencies, which convincingly demonstrate that pejorative labels (e.g., "bossy") appear more frequently in ratings of female instructors. However, notwithstanding the plethora of correlational studies, there is a need for natural experiments that tightly control all contextual factors that could be involved in the results reported, something that has only rarely been done.

In the domain of salary, taking together our own analysis of the NSF data and D. Li and Koedel's (2017) analysis, which controlled for productivity differences, we conclude that there is a gender salary gap of less than 4% for similar scientists, which is 80% smaller than the often-claimed 18.4% gap noted by the AAUP and others (e.g., Shen, 2013). Even this 4% unexplained gender salary gap may be due to factors such as career discontinuities, less-aggressive salary negotiations, gender differences in seeking competing offers, and grant awards—and, importantly, may also be due to women's lower overall productivity—rather than to overt gender bias (Perna, 2001). On the other hand, universities do have the choice not to respond to aggressive salary negotiations by men or to competing offers or to readjust the salaries of comparable female faculty when they do. In other words, if supply and demand in academic labor markets are creating these salary inequities on the basis of differential bargaining or mobility threats of men and women, even the economist on our team (S. K.) feels that the principle of "equal pay for truly equal work and ability" should, at some point, have a strong bearing on academic salary setting. Again, salary is a choice made by university administrators, and to the extent that there is a gender salary gap (and we did find evidence for such a gap), it can and should be addressed and eliminated.

Gender neutrality in four domains

In the other four domains, the evidence, although mixed, strongly favors the conclusion of gender neutrality in the areas of hiring for tenure-track jobs, grant funding in the United States, journal acceptances, and faculty recommendation letters. Regarding journal acceptances, we found in our meta-analysis some instances of male advantage but also some instances of female advantage, with most effects being very small and/or insignificant. Squazzoni et al.'s (2021) massive analysis, published 1 year outside the inclusion date for our meta-analysis, further reinforces the overall conclusion of gender neutrality. In the remaining three domains of hiring, U.S. grant funding, and recommendation-letter writing, the evidence from both dissection analysis and meta-analysis also points strongly to gender neutrality and, in the case of hiring, a pro-female advantage. However, in the case of grants, there was not gender neutrality in Europe or Canada, although the size of the male advantage did fall over the decades.

An important caveat to these conclusions is that they are restricted to tenure-track academia—we did not synthesize literature on nonprofessorial positions (e.g., postdocs, tech/lab workers, civil servants, and lecturers or in student-hiring simulations) or jobs in industry. There have been many demonstrations of gender bias in the evaluation of women in a variety of non-tenure-track contexts (e.g., Foschi et al., 1994; Koch et al., 2015; Lavy & Sand, 2015; Reuben et al., 2014; Swim et al., 1989). However, even in the realm of nonacademic hiring, Schaerer et al.'s (2022) large-scale meta-analysis and Birkelund et al.'s (2022) transnational analysis reported a pro-female hiring bias in recent years.

Note that the present analysis is in no way meant to deny the possibility of biases outside the six domains we studied or that bias existed in the past. Concerning the former, we have not here addressed tenure and promotion decisions (e.g., Ginther & Kahn, 2021; Weisshaar, 2017), professional awards (e.g., Cadwalader & Bryant-Friedrich, 2014; Van Miegroet & Glass, 2020), undercitation of papers by female authors (Mehto, 2021; Teich et al., 2022), invitations (e.g., Schroeder et al., 2014), "chilly climate," higher dropout rates from postdocs, sexual harassment, or persistence on the tenure track (e.g., Flaherty, 2018; Kaminski & Geisler, 2012; Settles et al., 2006).

Furthermore, this analysis does not address the potential derailment caused by deeper systemic forces, such as structural societal norms that impede women's progress. In fact, we argue below that the elimination of explicit bias in all six key areas of evaluation will enable scientists to turn their attention to addressing the important question of systemic sources of women's underrepresentation.

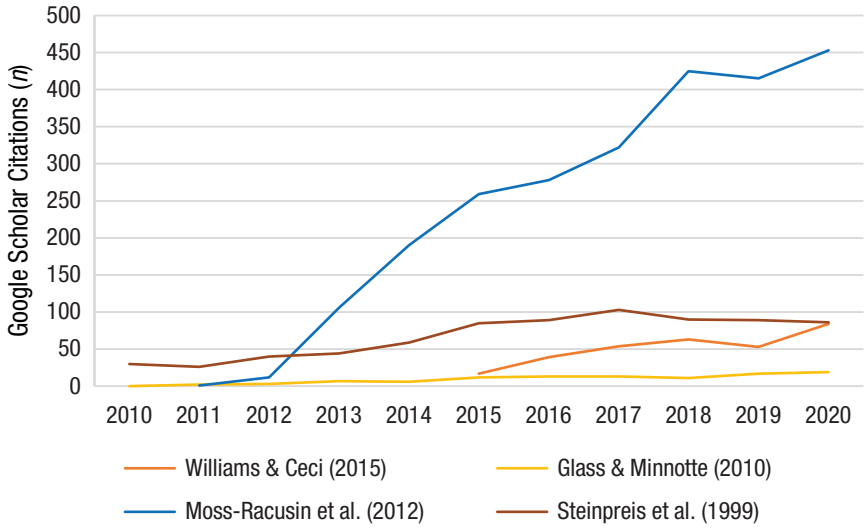

Fig. 7. Number of Google Scholar citations to four studies on academic hiring, from their year of publication to 2020. The two largest studies are cited least often.

Role of citation bias in fostering beliefs in existence of gender bias

Where do scientists' beliefs about gender bias originate? It appears that reliance on bias-confirming findings and ignoring of counterevidence has resulted in bias claims that do not always accord with the full corpus of findings and that sometimes are diametrically opposed to it (as appears to be the case in tenure-track hiring). Although we have not done careful bibliometric studies of this, our overall sense from scrutinizing the hundreds of articles cited here, as well as hundreds included in the meta-analyses we cited, is that many researchers who found bias in a given context also cited articles reporting bias in other contexts. For instance, an article reporting gender bias in grant awards tends to cite articles reporting gender bias in hiring, salary, and/or journal acceptances; an article reporting gender bias in salary tends to cite studies showing bias in other domains. Often, these authors do not cite articles that report no gender bias in the domain in question, let alone in other domains, and this may inflate a reader's sense of the magnitude of gender bias.

Consider the following: Despite broad and compelling evidence against the claim of bias from three types of tenure-track hiring studies (nationally representative cohort analyses, administrative records, and matched-CV experiments), most people's beliefs about gender bias in tenure-track hiring appear to be based on Moss-Racusin et al.'s (2012) famous study of faculty choosing between ambiguously competent BA-level applicants for a lab-manager job—not for a tenure-track faculty job. This study continues to be cited more than all contradictory tenure-track-focused studies combined (see Fig. 7). It is not obvious why Moss-Racusin et al. should be cited 7 times more often than the Williams and Ceci (2015) study from 2015 to 2020 (2,152 vs. 310 cites, a ratio that continues unabated today), because both studies were published in the same journal (*PNAS*), both were composed by mixed-gender author teams headed by a female first author, and both used the same experimental design.[28]

As another example of selective citation, Wennerås and Wold's (1997) report of gender bias in fellowship funding continues to be cited much more than the study that nullified their findings (and that even found some evidence of bias in favor of women), and it is also cited more than large metastudies (Fig. 8) that came to the opposite conclusion. Similarly, the earlier meta-analysis by Bornmann et al. (2007), which found signs of gender bias, continues to be cited at a higher rate than the study by many of the same authors (including Bornmann himself; Marsh et al., 2009) that improved Bornmann et al.'s statistical methodology and found no average bias—or even pro-female bias.

What could be responsible for such selective citation?

It is possible that systemic forms of gender bias not studied here—such as unequal expectations placed on young mothers (vs. fathers) struggling to succeed in tenure-track positions or cultural norms against women negotiating forcefully or performing in an agentic manner—may be in the back of people's minds when they think about the specific claims of bias examined here. The background awareness of such systemic

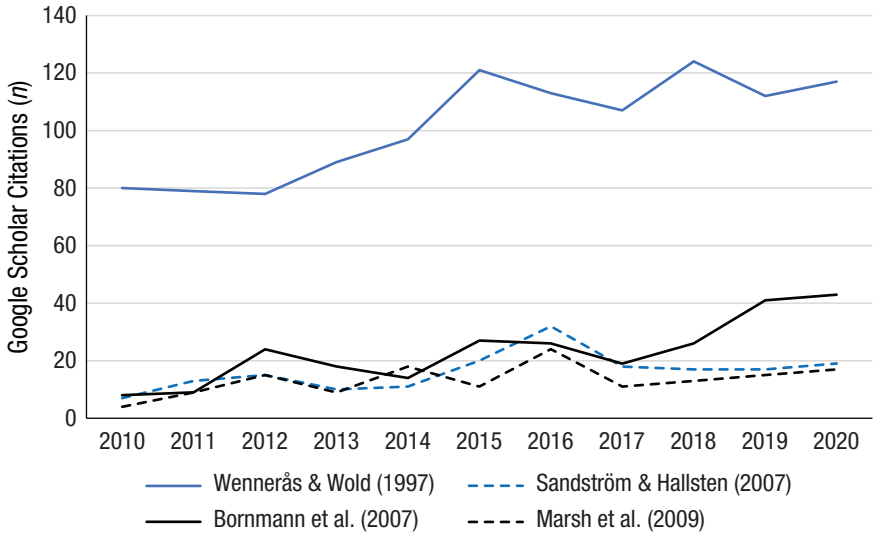

Fig. 8. Number of Google Scholar citations to four studies of grant funding, from their year of publication to 2020. The smallest and least generalizable study continues to be cited more than all of the larger contradictory ones combined.

factors may be responsible for creating unsupported beliefs about explicit bias in hiring, grant awarding, journal publishing, and letters of recommendation, and this could be one factor that partially drives the belief in and selective citation of bias claims noted throughout this article. People may believe that there is bias in one area, or they may have experienced bias themselves, and then they may generalize this thinking to an entire professional domain of academic science—even though the actual data may not support this generalization.

A number of other possible cognitive biases could be driving the selective citation illustrated in Figures 7 and 8. Some of these cognitive biases may be relevant to our findings of gender-neutral outcomes in four of the six domains and reduced bias in the other two domains. They may help us understand why some scientists may be aware of some gender-neutral evidence yet still believe in the widespread presence of explicit gender bias and be skeptical of claims of gender neutrality. And conversely, this same mechanism works both ways, such that other scientists may be so skeptical of evidence supporting gender bias that they downrate strong evidence for bias even when it does exist (see Handley et al., 2015, for relevant empirical evidence). In a working draft available from the authors (Ceci et al., 2023), we discuss five possible mechanisms that could be driving selective perception, appraisal, and citation: (a) ideological epistemology, (b) sociological networks, (c) storytelling, (d) editorial bias, and (e) moralistic fallacy. We do not offer these as a comprehensive list but merely as some possible mechanisms that could be driving the misjudgment of the scientific literature. Such misjudgment can lead to a false apparent consensus regarding controversial research by creating bias that leads to rejecting taboo conclusions (see Clark et al., 2023, and von Hippel & Buss, 2017, for empirical evidence that scientists' beliefs are in fact biased in this manner—see also Clark & Winegard, 2020).

Conclusion: Four Insights Derived From This Project

In sum, the full range of findings presented here provides no compelling evidence of widespread bias against women in four of six domains studied within academic tenure-track science (hiring, journal acceptances, U.S. grant funding, and letters of recommendation) and some evidence of bias in two domains (salary and teaching ratings), albeit with qualifications. We close by noting four insights derived from this 4.5-year adversarial collaboration. However, before doing so, we offer the following two caveats. The first is about the potential for implicit factors to create bias even in the absence of explicit factors, and the second is about bias that might have existed before our inclusion period began.

First, although explicit bias was not apparent in four of the six domains examined, it is important to note that implicit biases related to gendered expectations and stereotypes could nevertheless still lead to gendered outcomes, even in the absence of explicit bias. Social role theory (Eagly & Karau, 2002) posits gendered norms regarding behaviors to be rewarded versus sanctioned, such as the agentic–communal behavioral distinction described earlier. This concept could explain

the occupational segregation seen in Figure 2 because it posits that natural-science fields are less congruent with stereotypical female role expectations in comparison with social science and humanities fields. There is sometimes an implicit disjunction between being a woman and being a scientist (Heilman, 1983; Nosek et al., 2002; Smyth & Nosek, 2015). Despite the increases in women entering GEMP fields in recent decades, older children continue to associate science with being male (D. I. Miller et al., 2018). Social role theory also suggests that salary gaps could occur even in the absence of explicit discrimination, because engaging in vigorous negotiating over salary violates female gender-role expectations, which can lead to negative consequences for women. Similarly, some evidence shows that female instructors are downrated for agentic behavior, such as giving negative grading feedback (Buser et al., 2022). These are a few examples of the many possible roles of implicit factors in domains where explicit biases are not found.

Second, as noted above, our findings of some areas of gender neutrality or even a pro-female advantage are very much rooted in the most recent decades and in no way minimize or deny the existence of gender bias in the past. Throughout this article, we have noted pre-2000 analyses that suggested that bias either definitely or probably was present in some aspects of tenure-track academia before 2000. This was particularly true for grants and the salary gap. An example comes from a context we did not study, a recent study by Card and his colleagues (2022) documenting temporal trends in prestigious awards. Between 1960 and 1990, women had a lower chance of being inducted into the National Academies of Science and the American Academy of Arts and Sciences, but this disadvantage became gender neutral around 1990, and starting around 2000, women became 3 to 15 times more likely to be inducted into the these organizations than men with comparable publications and citations. Such analyses remind us that even if the academic landscape today is often gender neutral or preferential for women, it was not always so.

Such improvements might have occurred because of earlier reports noting the existence of bias. Granting agencies in the United States are particularly conscious of the need to avoid any hint of unfair behavior toward women and minorities. The same is likely true of some journal editors and university administrators who audit salaries. Similarly, feminist responses to the underrepresentation of women in GEMP, which may have prodded GEMP departments to do better, may have led to these departments trying harder to recruit women over the last two decades, with (for example) search committee members being increasingly required to take diversity training courses prior to serving on hiring committees.

Insight 1

There are reasons to believe that in at least three of the domains we studied, the findings of gender neutrality may be due to the qualifications and work products of job applicants, grant PIs, and journal authors being made explicit to evaluators, who were often highly motivated to make the right choice. Evidence from a large meta-analysis of gender bias shows that evaluation results differ as a function of the professional motivation of evaluators and how explicit applicants' high level of competence is made to evaluators (Koch et al., 2015). These two factors are relevant to tenure-track evaluations, given the motivation of tenure-track search committees and the excellence of short-listed applicants for tenure-track positions (see Note 9 for support from a national canvas for the claim that short-listed applicants for tenure-track jobs are judged to be excellent) and the fact that hiring committees have access to CVs, job talks, meetings, and letters of recommendation.

Such unequivocal indicators of excellence are rarely available when hiring outside academia, although recent large-scale analyses report that even in nonacademic domains, there has been a trend in the past two decades for hiring to favor women (Birkelund et al., 2022; Schaerer et al., 2022). These indicators are also not available when students choose to rate teachers. This suggests that the place to search for gender bias might be in contexts in which people have only ambiguous or equivocal indicators of competence, which are not those typically found for tenure-track short lists. At the same time, this suggests that in order to prevent bias in any evaluative process—be it hiring, grant making, or reviewing—having clear criteria regarding what one is looking for is imperative, as are ample opportunities for all applicants to provide detailed unambiguous information and evidence regarding their competence.

Insight 2

It is useful to identify domains without current explicit gender bias, for three reasons. First, it encourages new efforts pinpointed at what could be currently causing inequities and underrepresentation of women. For instance, if a key issue is that women's careers are undermined by underlying systemic factors—such as women not applying for tenure-track jobs or grants because they are overextended with juggling family care and academic work, as Xu (2015) showed is the case for women pursuing nonacademic positions in STEM fields—perhaps we need to think more deeply about how academia can be made more flexible in its timing of milestones (particularly in fields requiring

long postdocs). If student-teaching evaluations are gender biased, perhaps they should be supplemented by objectively evaluating the actual learning that students have mastered. If men have higher salaries because they generate competing offers or are more comfortable negotiating, perhaps universities need to have regular audits of salary gaps and raise salaries for faculty whose pay is inequitable, precisely because they have not sought outside offers or negotiated for higher wages.

Second, institutions should get credit for progress they have made as a result of decades of efforts and be encouraged to continue these efforts while making progress on new fronts. For instance, U.S. funding agencies should be lauded for their fair review of women's and men's new grants. However, they need to also increase their efforts to encourage women to apply for new grants, to resubmit grants that were turned down, and to apply for grant continuations. Similarly, universities should continue gender-fair processes of evaluations of applications but perhaps also encourage more women to apply by introducing flexibility in employment.

Third, if women believe that every step of academia is biased against them, some may be reluctant to enter academia. They may instead seek industry or government jobs, which are often the location of much creative scientific work. Or women may become health professionals but not health researchers. As a result, women's representation among the individuals training new generations will remain lower than it could and should be.

Insight 3

We all have to do a better job of verifying facts. Despite our finding of little bias against women in four of the domains we examined (and a pro-female bias in tenure-track hiring), many scholars and advocates appear to believe, state, and write otherwise. As seen in the quotes and committee statements throughout this article, prior to our analyses, many scholars claimed that women were unquestionably penalized in all six areas we examined, and these claims were found in top outlets such as *Science*, *Nature*, *PNAS*, *The New York Times*, and NASEM consensus reports. In support of their beliefs, the authors selectively cited congruent evidence and ignored contrary findings, which they may hold to a higher standard of evidence because of their prior beliefs.

Elsewhere, we posit potential mechanisms that may be responsible for the failure to appreciate the totality of evidence (contact the corresponding author for details). It is essential to base conclusions on the full range of evidence, filtered for methodological limitations, and acknowledge boundary conditions rather than make overarching claims. This point was recently made by researchers studying hiring in philosophy, in response to the claim by some of implicit bias against women in their field. They in fact found significant pro-female hiring bias, prompting them to argue, "Factors such as implicit bias should give women a disadvantage in the academic job market, which, again, is not what our analysis shows" (Kallens et al., 2022, p. 673).

Insight 4

Editors and board members can promote science by encouraging, when possible, diverse viewpoints and by commissioning teams of adversarial coauthors (as this particular journal, *Psychological Science in the Public Interest*, was founded to do—to bring coauthors together in an attempt to resolve their historic differences). Knowing that one's writing will be criticized by one's divergently thinking coauthors can reduce ideologically driven criticisms that are offered in the guise of science. Unlike open-science initiatives that maximize replicability and reduce *p*-hacking and hypothesizing after the results are known (HARKing)—such as preregistration of hypotheses, specification of sample size and planned statistical analyses—adversarial collaborations provide a missing element that open science was not designed to prevent. Open-science preregistration is usually conducted by like-minded scientists, and hence it does not prevent researchers from cherry-picking their methods or operationalizing definitions to make it easier to support their hypotheses, which can lead them to design collaborations that tilt the outcome toward confirmation of their hypotheses. In contrast,

> adversarial collaborators have to negotiate among themselves the framing of hypotheses, the operationalization of constructs and definitions, the most suitable methods to use, and what outcome each member of an adversarial team will accept as evidence against its position. None of these considerations are inherent in preregistration. (Ceci & Williams, 2022, p. 35)

Failure to ensure meaningful viewpoint diversity among team members can lead to major misunderstandings of the corpus of scientific evidence, with potential scholarly as well as real-world costs. As our own adversarial collaboration has taught us, we all need to remain open-minded regarding alternative views rather than prematurely assume that the science is settled.

Transparency

Editor: Nora S. Newcombe

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Stephen J. Ceci https://orcid.org/0000-0002-2835-9707

Notes

1. Searching "women or female or gender" and "STEM or science or engineering or math" and "faculty or academ* or universit*" yields 17,100 new papers or books in Google Scholar from mid-2013 to 2020. Using a different methodology, Elsevier's (2017) massive analysis of gender research across 27 countries concluded the growth in published papers has been even greater: "Between 2011 and 2015, there are over 23,000 gender research papers, more than two and a half times (2.7 factor of growth) the number of [gender] papers published between 1996 and 2000" (p. 71).

2. Witteman et al. (2019), to their credit, also list areas in which women are not disadvantaged.

3. As noted in the preface, our research team was initially constituted in 2012 as part of an adversarial collaboration by the *Psychological Science in the Public Interest* editorial board to write our former joint review article (Ceci et al., 2014). The team members did not know each other, but we knew of each other's published work. We came together explicitly because we had different views on the debate over gender bias in academic science; we had published very different claims in the past and saw this collaboration as an opportunity to see whether we could find common ground. Rather than automatically ratifying each other's claims, we challenged any claim that a team member felt was based on a misreading of the evidence or the result of unrepresentative sampling. This viewpoint difference among team members helped reduce the possibility that literature searches could be countered by studies not cited.

4. All numbers are from the U.S. National Center for Education Statistics (Snyder et al., 2019). The historical data are from Snyder (1993).

5. We examined an eighth report, and its findings accord with those we review here. However, we deleted this eighth report because its authors did not wish it to be made public and thus there is no way for readers to confirm our analysis.

6. This might be viewed as desirable favoritism to offset historical bias in the French academy against women and a success story for earlier findings of gender bias that led to such favoritism.

7. In spring 2020, we conducted a national random survey of tenure-track faculty in seven scientific fields plus philosophy, asking them to rate on a 10-point scale the competence of the top three applicants in their most recent tenure-track search. Two hundred seventy-nine tenure-track faculty responded and their median rating was 9.0 (Quarter 3–Quarter 1 = 2.55, indicating clustering at the right tail); 88.5% of respondents said that the top candidate in their search was between *very strong* and *truly outstanding*. Even their top three applicants were rated, on average, as *excellent*. This appears to be a reality in tenure-track hiring: Applicants who make it to the short list are seen as excellent, which may not be the case at lower levels, such as postdocs, lab managers, and student employees.

8. A number of similar findings appeared after our 2020 termination date (Henningsen et al., 2021; Kallens et al., 2022; Solga et al., 2023). Henningsen and her associates published their data several months after the close of the inclusion period, so we did not include their data in our analysis, but had we done so, the conclusions we reached in this article would be further bolstered. These researchers asked 481 economists at German (76.7%), Swiss (15.6%), and Austrian (4.4%) universities to assume the role of a search committee member. Using a procedure similar to that of Williams and Ceci (2015), the authors asked participants to rank three short-listed applicants' competence and hirability for a tenure-track position. Both male and female economists rated the woman as 3 times more hirable than the identically accomplished male counterpart: "The female candidate was ranked overall as the first choice for the professorship at an approximate ratio of 3:1" (p. 43). Similarly, Solga et al.'s (2023) recent experimental study of 1,688 German professors also found a modest female advantage in both perceived competence/qualification and in being invited, and this was true of both male and female professors serving on a hypothetical search committee. Their findings mimicked the pro-female advantage in national appointment data in Germany. Kallens et al. (2022) reported a significant pro-female hiring advantage in philosophy, which coupled with earlier findings in that field (Allen-Hermanson, 2017), suggest a 10% to 17% advantage.

9. The largest meta-analysis of field audits by Schaerer et al. (2022) spanned 44 years of findings (1976–2020) and 373,706 applicants from 87 field audits (240 effect sizes), finding that "contrary to the expectations of both laypeople and scientists, who believed that hiring discrimination remains widespread, selection bias in favor of male over female candidates was eliminated and, if anything, slightly reversed in sign starting in 2009" (p. 1). For similar findings, see Birkelund et al. (2022).

10. In this section, we omit studies of grants that did not estimate gender differences in success of grant applications. This includes Chandler (2018), Jagsi et al. (2009), Kaatz et al. (2015), Lerchenmueller and Sorenson (2018), Magua et al. (2017), and Warner et al. (2017), among others. These and other articles not included in our meta-analysis instead measure such things as whether faculty members have received a grant by a certain career stage and textual analysis of reviews.

11. Search terms were "women or gender or female or bias or sex" and "grant* or proposal* or fund*."

12. We also estimated odds ratios and relative risk ratios. All qualitative results were the same as reported for Cohen's *d* using meta regression and using regular regression weighting and clustering.

13. This multivariate analysis was performed both on meta regression and using regular regression, weighting by number of applications and clustering by article.

14. Many scholars believe that PIs who have been successful in past research and publications are more likely to be successful in their currently proposed research, a Matthew effect that

results in a nonrandom cumulative advantage (Lawson et al., 2021; for broader claims of the validity of cumulative advantage, see Perc, 2014). In the past, many grant agencies agreed and instructed reviewers to consider the publication record of PIs.

15. Fox (2005) reported that because of their time allocation, women with preschool children were actually more productive in publishing than were women with school-age children and women with no children.

16. Using a 6-point scale, Rivera and Tilcsik (2019) found that there was no gender gap in rating courses taught by a woman (0.02 points lower, $SE = 0.14$, n.s.), whereas women were rated 0.52 points lower when using a 10-point scale ($SE = 0.28$, $p < .10$).

17. Schmidt's (2020) interactive site allows users to see the frequency of words in 14 million RMP reviews. Nearly all negative words ("unkind," "cruel," "bossy," "mean," "disorganized") are more frequent in comments made about female instructors. In contrast, words such as "genius" were used more often to describe male instructors (Storage et al., 2016).

18. Although they published their study a few months after the end of our inclusion date, Wong and Bouchard (2021) reported a similar result for 232 students followed over a multisemester online course. Students rated teaching assistants (TAs) who were presented as male as having higher expectations and their courses as being more challenging than TAs who were presented as female.

19. In natural sciences and most social sciences, the first author is typically one who has made the largest contribution to the article and the last author is the lab head or senior author. The main exception is economics (and law reviews), where authors are alphabetical. There is less agreement about who corresponding authors should be. Biased reviewers may implicitly identify a paper's gender by one or more of these three author roles.

20. This claim has repeatedly been called into doubt (Gelman, 2019; Pallesen, 2019; Sommers, 2020).

21. Two studies fell outside the inclusionary boundary (2000–2020). One was a high-profile analysis by the economists Ding et al. (2021) that appeared in *Nature Biotechnology*, and some readers may wonder whether its findings would alter the conclusions from the 2000 to 2020 studies. Ding et al. controlled for a long list of factors in their analysis of gendered pay gaps in industry and academia—both tenure track versus nontenure track. It generated substantial visibility and press attention. These economists reported that a significant gender salary gap exists, but it is confined to non-tenure-track faculty. Tenured and tenure-track positions showed gender neutrality, reinforcing our conclusion. The other was an extensive analysis of 17 R1 universities by Samaniego et al. (2023) that reported that women and men were remunerated equivalently for increases in their *h* indices in the social and behavioral sciences, but women were underremunerated vis-à-vis men in the STEM fields; 22.47% of the salary variance was accounted for by disciplinary differences, though this still left pay inequity in favor of men when STEM fields were separated from social and behavioral science fields.

22. Self-reported weekly hours worked is a less than ideal measure, especially when it comes to gender, as some scholars argue that men overestimate and women underestimate. If true, this raises caution about other self-reported data, such as those on chilly climate and harassment.

23. It is known, however, that in many contexts, women are judged more harshly than men when they attempt to establish their authority the same way that men do (Eagly et al., 2003). University students and college-educated adults both expect women to fulfill prescriptions of feminine niceness, such as sensitivity to the needs of others and lack of pushiness, neither of which are compatible with the prototypical male style of hard negotiating and refusal to accept low salary offers (Bowles et al., 2007). In their experiments, Bowles and her colleagues found that women were penalized more than men, especially when male participants were evaluators. Subsequent meta-analyses by Kugler and her associates (Kugler et al., 2018; Reif et al., 2019) demonstrated that gender differences in initiating negotiation were small and context dependent and were even smaller when the situational ambiguity of negotiating was low. Interestingly, in their meta-analysis, Kugler et al. reported that there was no relationship between willingness to initiate negotiation and a country's level of gender inequality.

24. Madera et al. (2009) was archived online in 2020, hence searchable within our parameters.

25. We included this study even though it involved postdocs rather than tenure-track applicants because it is high profile (published in *Nature* with extensive media attention) and to prevent readers from wondering whether its inclusion would change the overall conclusions. Excluding it would strengthen the conclusion.

26. What follows is not a comprehensive review of the enormous literature on productivity but instead a review of reviews based on the largest, most exhaustive analyses that we are aware of, based on our extensive reading and writing on this topic (e.g., Elsevier, 2017; Huang et al., 2020; Larivière et al., 2013), as well as citations by other researchers to these large studies. Almost without exception, this literature points in the same direction, including the myriad reviews we cite.

27. West et al. (2013) also published another large bibliometric analysis of more than 8 million articles in the JSTOR network data set. They looked at time patterns in the numbers of female authors from 1900 to 2010. They found that women made progress in being first authors over the period from 1990 to 2010. But their emphasis was representation among authors, which picks up both trends in the number of women academics and the number of publications per person.

28. Specifically, Williams and Ceci's (2015) fifth experiment was conducted to mimic Moss-Racusin et al.'s (2012) experimental design. The political scientist Zigerell (2018) came to a similar conclusion, adjusting for the number of years since publication. For example, he shows that the Williams and Ceci article had 15 Web of Science citations 3 years after its publication, whereas the Moss-Racusin et al. article had 109, a trend that continues a decade after publication. (For other years, see Zigerell, 2021; the social psychologist Jussim, 2019, also arrived at a similar conclusion).

References

Abel, M. H., & Meltzer, A. L. (2007). Student ratings of a male and female professors' lecture on sex discrimination in the workforce. *Sex Roles*, *57*(3–4), 173–180.

Abramo, G., Aksnes, D. W., & D'Angelo, C. A. (2021). Gender differences in research performance within and between countries: Italy vs Norway. *Journal of Informetrics*, *15*, 101144. https://doi.org/10.1016/j.joi.2021.101144

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2007). Gender differences in research productivity: A bibliometric analysis of the Italian academic system. *Scientometrics*, *79*(3), 517–539.

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). The contribution of star scientists to overall sex differences in research productivity. *Scientometrics*, *81*(1), 137–156.

Aguinis, H., Ji, Y. H., & Joo, H. (2018). Gender productivity gap among star performers in STEM and other scientific fields. *Journal of Applied Psychology*, *103*, 1283–1306. https://doi.org/10.1037/apl0000331

Aiston, S. J., & Jung, J. (2015). Women academics and research productivity: An international comparison. *Gender and Education*, *27*(3), 205–220.

Albers, C. J. (2015). Dutch research funding, gender bias, and Simpson's paradox. *Proceedings of the National Academy of Sciences, USA*, *112*, E6828–E6829.

Allen-Hermanson, S. (2017). Leaky pipeline myths: In search of gender effects on the job market and early career publishing in philosophy. *Frontiers in Psychology*, *8*, Article 953. https://doi.org/10.3389/fpsyg.2017.00953

Alvarez, S. N. E., Jagsi, R., Abbuhl, S. B., Lee, C. J., & Myers, E. R. (2019). Promoting gender equity in grant making: What can a funder do? *The Lancet*, *393*(10171), E9–E11. https://doi.org/10.1016/S0140-6736(19)30211-9

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, *12*(5), 385–390. https://doi.org/10.1111/1467-9280.00371

American Association of University Professors. (2019). *The annual report on the economic status of the profession, 2018–19*. https://www.aaup.org/sites/default/files/2018-19_ARES_Final_0.pdf

American Psychological Association, Committee on Women in Psychology. (2017). *The changing gender composition of psychology: Update and expansion of the 1995 task force report*. American Psychological Association. https://www.apa.org/pi/women/programs/gender-composition/task-force-report.pdf

Appel-Cresswell, S., Blanchet, P. J., Wysocki, J., & Postuma, R. B. (2019). Gender and funding success. *The Lancet*, *393*(10187), 2195–2196.

Astegiano, J., Sebastián-González, E., & Castanho, C. (2019). Unravelling the gender productivity gap in science: A meta-analytical review. *Royal Society Open Science*, *6*(6), Article 181566. https://doi.org/10.1098/rsos.181566

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education*, *48*, 193–210.

Basow, S., Codos, S., & Martin, J. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal*, *47*(2), 352–363.

Bataille, P., Le Feuvre, N., & Kradolfer Morales, S. (2017). Should I stay or should I go? The effects of precariousness on the gendered career aspirations of postdocs in Switzerland. *European Educational Research Journal*, *16*, 313–331. https://doi.org/10.1177/1474904116673372

Bautista-Puig, N., García-Zorita, C., & Mauleón, E. (2019). European Research Council: Excellence and leadership over time from a gender perspective. *Research Evaluation*, *28*, 370–382. https://doi.org/10.1093/reseval/rvz023

Berg, J. (2017). Looking inward at gender issues. *Science*, *355*(6323), Article 329. https://doi.org/10.1126/science.aam8109

Berg, J. (2019, January 3). New tools for gender analysis. *Sciencebound*. https://blogs.sciencemag.org/sciencehound/2019/01/03/new-tools-for-gender-analysis/

Bernard, C. (2018). Editorial: Gender bias in publishing: Double-blind reviewing as a solution? *eNeuro*, *5*(3), Article e0225-18.2018. https://doi.org/10.1523/ENEURO.0225-18.2018

Bernstein, R. H., Macy, M. W., Cameron, C., Williams-Ceci, S., Williams, W. M., & Ceci, S. J. (2022). Assessing gender bias in particle physics and social-science recommendations for academic jobs. *Social Sciences*, *11*, Article 74. https://doi.org/10.3390/socsci11020074

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*, 389–391. https://doi.org/10.1126/science.aah6524

Bian, L., Leslie, S.-J., & Cimpian, A. (2018). Evidence of bias against girls and women in contexts that emphasize intellectual ability. *American Psychologist*, *73*(9), 1139–1153.

Biernat, M. (2012). Stereotypes and shifting standards: Forming, communicating, and translating person impressions. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 45, pp. 1–59). Academic Press.

Biernat, M., & Eidelman, S. (2007). Translating subjective language in letters of recommendation: The case of the sexist professor. *European Journal of Social Psychology*, *37*(6), 1149–1175. https://doi.org/10.1002/ejsp/432

Bird, K. S. (2011). Do women publish fewer journal articles than men? Sex differences in publication productivity in the social sciences. *British Journal of Sociology of Education*, *32*(6), 921–937. https://doi.org/10.1080/01425692.2011.596387

Birkelund, G. E., Lancee, B., Larsen, E. N., Polavieja, J. G., Radl, J., & Yemane, R. (2022). Gender discrimination in hiring: Evidence from a cross-national harmonized field experiment. *European Sociological Review*, *38*(3), 337–354. https://doi.org/10.1093/esr/jcab043

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from The *American Economic Review*. The *American Economic Review*, *81*(5), 1041–1067.

Blinkenstaff, J. C. (2005). Women and science careers: Leaky pipeline or gender filter? *Gender and Education*, *17*, 369–386.

Blue, J., Traxler, A. L., & Cid, X. C. (2018). Gender matters. *Physics Today*, *71*(3), 40–46. https://doi.org/10.1063/PT.3.3870

Bordons, M., Morillo, F., Fernández, M. T., & Gómez, I. (2003). One step further in the production of bibliometric indicators at the micro level: Differences by gender and professional category of scientists. *Scientometrics*, *57*(2), 159–173.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *45*, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Bornmann, L., & Daniel, H. (2005). Selection of research fellowship recipients by committee peer review. *Scientometrics*, *63*, 297–320.

Bornmann, L., Mutz, R., & Daniel, H. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226–238.

Borsuk, R. M., Aarssen, L. W., Budden, A. E., Koricheva, J., Leimu, R., Tregenza, T., & Lortie, C. J. (2009). To name or not to name: The effect of changing author gender on peer review. *BioScience*, *59*(11), 985–989. https://doi.org/10.1525/bio.2009.59.11.10

Bowles, H. R., Babcock, L., & Lai, L. (2007). Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does not hurt to ask. *Organizational Behavior and Human Decision Processes*, *103*, 84–103. https://doi.org/10.1016/j.obhdp.2006.09.00

Braisher, T. L., Symonds, M. R. E., & Gemmell, N. J. (2005). Publication success in *Nature* and *Science* is not gender dependent. *BioEssays*, *27*(8), 858–859. https://doi.org/10.1002/bies.20272

Breda, T., & Hillion, M. (2016). Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science*, *353*(6298), 474–478.

Broder, I. E. (1993). Review of NSF economics proposals: Gender and institutional patterns. *The American Economic Review*, *83*(4), 964–970.

Brooks, J., & Della Sala, S. (2009). Re-addressing gender bias in Cortex publications. *Cortex*, *45*(10), 1126–1137. https://doi.org/10.1016/j.cortex.2009.04.004

Brower, A., & James, A. (2020). Research performance and age explain less than half of the gender pay gap in New Zealand universities. *PLOS ONE*, *15*(1), Article e0226392. https://doi.org/10.1371/journal.pone.0226392

Budden, A. E., Tregenza, T., Aarssen, L. W., Koricheva, J., Leimu, R., & Lortie, C. J. (2008). Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, *23*(1), 4–6. https://doi.org/10.1016/j.tree.2007.07.008

Burns, K. E. A., Straus, S. E., Liu, K., Rizvi, L., & Guyatt, G. (2019). Gender differences in grant and personnel award funding rates at the Canadian Institutes of Health Research based on research content area: A retrospective analysis. *PLOS Medicine*, *16*(10), Article e1002935. https://doi.org/10.1371/journal.pmed.1002935

Buser, W., Batz-Barbarich, C. L., & Hayter, J. K. (2022). Evaluation of women in economics: Evidence of gender bias following behavioral role violations. *Sex Roles*, *86*, 695–710.

Cadwalader, E. L., & Bryant-Friedrich, A. C. (2014). Improving transparency in scholarly recognition by scientific societies. In H. N. Cheng, S. Shah, & M. L. Wu (Eds.), *Careers, entrepreneurships, and diversity: Challenges and opportunities in the global chemistry enterprise: American Chemical Society Series* (Vol. 1169, pp. 245–254). https://doi.org/10.1021/bk-2014-1169.ch022

California State Auditor, Bureau of State Audits. (2001, May). *University of California: Some campuses and academic departments need to take additional steps to resolve gender disparities among professors*. https://www.bsa.ca.gov/pdfs/reports/2000-131.pdf

Cameron, E. Z., White, A. W., & Gray, M. E. (2016). Solving the productivity and impact puzzle: Do men outperform women, or are metrics biased? *BioScience*, *66*(3), 245–252. https://doi.org/10.1093/biosci/biv173

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, *135*(1), 269–327.

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2022). *Gender gaps at the academies* (NBER Working Paper No. 30510). National Bureau of Economic Research. https://doi.org/10.3386/w30510

Carey, J. M., Carman, K. R., Klayton, K. P., Horiuchi, Y., Htun, M., & Ortiz, B. (2020). Who wants to hire a more diverse faculty? A conjoint analysis of faculty and student preferences for gender and racial/ethnic diversity. *Politics, Groups and Identities*, *8*, 535–553.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, *134*(3), 1163–1224.

Carlana, M., & Corno, L. (2021). *Parents and peers: Gender stereotypes in the field of study* (CEPR Discussion Paper No. DP16582). https://ssrn.com/abstract=3960154

Carlsson, F., Löfgren, Å., & Sterner, T. (2012). Discrimination in scientific review: A natural field experiment on blind versus non-blind reviews. *The Scandinavian Journal of Economics*, *114*(2), 500–519.

Carlsson, M., Finseraas, H., Midtbøen, A. H., & Rafnsdóttir, G. L. (2021). Gender bias in academic recruitment? Evidence from a survey experiment in the Nordic region. *European Sociological Review*, *37*(3), 399–410. https://doi.org/10.1093/esr/jcaa050

Cassad, B. J., Franks, J. E., Garasky, C. E., Kittleman, M. M., Roesler, A. C., Hall, D. Y., & Petzel, Z. W. (2021). Gender inequality in academia: Problems and solutions for women faculty in STEM. *Journal of Neuroscience Research*, *99*(1), 12–23. https://doi.org/10.1002/jnr.24631

Casselman, B. (2021, February 23). For women in economics, the hostility is out in the open. *The New York Times*. https://www.nytimes.com/2021/02/23/business/economy/economics-women-gender-bias.html

Cech, E. A., & Blair-Loy, M. (2019). The changing career trajectories of new parents in STEM. *Proceedings of the National Academy of Sciences, USA*, *116*, 4182–4187.

Ceci, S. J. (2018). Women in academic science: Experimental findings from hiring studies. *Educational Psychologist*, *53*(1), 22–41.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*(3), 75–141. https://doi.org/10.1177/1529100614541236

Ceci, S. J., Kahn, S., & Williams, W. M. (2023). *Repository A* [Unpublished manuscript]. Department of Psychology, Cornell University.

Ceci, S. J., & Peters, D. P. (1984). How blind is blind review? *American Psychologist*, *39*, 1491–1494.

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences, USA*, *108*(8), 3157–3162. https://doi.org/10.1073/pnas.1014871108

Ceci, S. J., & Williams, W. M. (2015). Women have substantial advantage in STEM faculty hiring, except when competing against more-accomplished men. *Frontiers in Psychology*, *6*, Article 1532. https://doi.org/10.3389/fpsyg.2015.01532

Ceci, S. J., & Williams, W. M. (2022). The importance of viewpoint diversity among scientific team members: Comment on Clark et al. *Journal of Applied Research on Memory and Cognition*, *11*(1), 35–40. https://doi.org/10.1037/mac0000007

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, *71*, 17–33. https://doi.org/10.1080/00221546.2000.11780814

Chandler, V. (2018). Do evaluators prefer candidates of their own gender? *Canadian Public Policy*, *44*(4), 289–302.

Chatman, J. A., Sharps, D., Mishra, S., Kray, L. J., & North, M. S. (2022). Agentic but not warm: Age-gender interactions and the consequences of stereotype incongruity perceptions for middle-aged professional women. *Organizational Behavior and Human Decision Processes*, *173*, Article 104190. https://doi.org/10.1016/j.obhdp.2022.104190

Cheng, S. D. (2020). *Careers versus children: How childcare affects academic tenure-track gender gap*. https://scholar.harvard.edu/files/sdcheng/files/sdcheng_jmp.pdf

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1–35.

Chevalier, J. (2019). *The 2019 report of the Committee on the Status of Women in the Economics Profession*. American Economic Association. https://www.aeaweb.org/content/file?id=11672

Clark, C. J., Fjeldmark, M., Lu, L., Baumeister, R. F., Ceci, S., German, K., Reilly, W., Tice, D., von Hippel, W., Williams, W., Winegard, B. M., & Tetlock, P. E. (2023, February 24). *Taboos and self-censorship among psychology professors* [Conference session]. Society for Open Inquiry in Behavioral Science, Behavioral Science Speakeasy, Atlanta, GA, United States.

Clark, C. J., & Winegard, B. M. (2020). Tribalism in war and peace: The nature and evolution of ideological epistemology and its significance for modern social science. *Psychological Inquiry*, *31*(1), 1–22. https://doi.org/10.1080/1047840X.2020.1721233

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, *1*(1), Article e1400005. https://doi.org/10.1126/sciadv.1400005

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Routledge Academic.

Coil, A. (2017, August 25). Why men don't believe the data on gender bias in science. *Wired*. https://www.wired.com/story/why-men-dont-believe-the-data-on-gender-bias-in-science/

Cole, J. R., & Zuckerman, H. (1984). The productivity puzzle: Persistence and change in patterns of publications of men and women scientists. In M. W. Steinkamp & M. L. Maehr (Eds.), *Advances in motivation and achievement: Women in science* (pp. 217–258). JAI Press.

Cuskley, C., Roberts, S. G., Politzer-Ahles, S., & Verhoef, T. (2020). Double-blind reviewing and gender biases at EvoLang conferences: An update. *Journal of Language Evolution*, *5*(1), 92–99.

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development*, *82*, 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.xpmid:21410915

D'Amico, R., Vermigli, P., & Canetto, S. (2011). Publication productivity and career advancement by female and male psychology faculty: The case of Italy. *Journal of Diversity in Higher Education*, *4*, 175–184. https://doi.org/10.1037/a0022570

Davis, G. (2009). Improving the postdoctoral experience: An empirical approach. In G. Freeman (Eds.), *Science and engineering careers in the United States: An analysis of markets and employment* (pp. 99–127). University of Chicago Press.

Dehdarirad, T., Villarroya, A., & Barrios, M. (2015). Research on women in science and higher education: A bibliometric analysis. *Scientometrics*, *103*(3), 795–812.

Ding, W. V., Ohyama, A., & Agarwal, R. (2021). Trends in gender pay gaps of scientists and engineers in academia and industry. *Nature Biotechnology*, *39*, 1019–1026. https://doi.org/10.1038/s41587-021-01008-0

Doubleday, A. F., & Lee, L. M. J. (2016). Dissecting the voice: Health professions students' perceptions of instructor age and agender in an online environment and the impact on evaluations for faculty. *Anatomical Sciences Education*, *9*(6), 537–544. https://doi.org/10.1002/ase.1609

Duch, J., Zeng, X. H. T., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K., & Amaral, L. A. N. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLOS ONE*, *7*(12), Article e51332. https://doi.org/10.1371/journal.pone.0051332

Dunham, Y., Baron, A. S., & Banaji, M. R. (2015). The development of implicit gender attitudes. *Developmental Science*, *19*, 781–789. https://doi.org/10.1111/desc.12321pmid:26260250

Dutt, K., Pfaff, D. L., Bernstein, A. F., Dillard, J. S., & Block, C. J. (2016). Gender differences in recommendation

letters for postdoctoral fellowships in geoscience. *Nature Geoscience*, *9*(11), 805–808. https://doi.org/10.1038/NGEO2819

Eagly, A. H. (2020). Do the social roles that women and men occupy in science allow equal access to publication? *Proceedings of the National Academy of Sciences, USA*, *117*(11), 5553–5555. https://doi.org/10.1073/pnas.2001684117

Eagly, A. H., Johannesen-Schmidt, M. C., & van Engen, M. L. (2003). Transformational, transactional and laissez-faire leadership styles: A meta-analysis comparing men and women. *Psychological Bulletin*, *129*, 569–591.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 123–174). Erlbaum.

Eaton, A. A., Saunders, J. F., Jacobson, R. K., & West, K. (2020). How gender and race stereotypes impact the advancement of scholars in STEM: Professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, *82*, 127–141.

Ecklund, E. H., & Lincoln, A. E. (2011). Scientists want more children. *PLOS ONE*, *6*(8), Article e22590. https://doi.org/10.1371/journal.pone.0022590

Elsevier. (2017). *Gender in the global research landscape*. https://www.elsevier.com/__data/assets/pdf_file/0008/265661/ElsevierGenderReport_final_for-web.pdf

European Molecular Biology Organization. (2019). *EMBO facts & figures*. https://www.embo.org/documents/news/facts_figures/EMBO_facts_figures_2019.pdf

Engqvist, L., & Frommen, J. G. (2008). Double-blind peer review and gender publication bias. *Animal Behavior*, *76*(3), e1–e2. https://doi.org/10.1016/j.anbehav.2008.05.023

European Commission, Directorate-General for Research and Innovation. (2019). *She figures 2018*. https://data.europa.eu/doi/10.2777/936

Fan, Y., Shepherd, L., Slavich, D., Waters, D., Stone, M., Abel, R., & Johnston, E. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLOS ONE*, *14*(2), Article e0209749. https://doi.org/10.1371/journal.pone.0209749

Farr, C. M., Bombaci, S. P., Gallo, T., Mangan, A. M., Riedl, H. L., Stinson, L. T., Wilkins, K., Bennett, D. E., Nogeire-McRae, T., & Pejchar, L. (2017). Addressing the gender gap in distinguished speakers at professional ecology conferences. *BioScience*, *67*(5), 464–468. https://doi.org/10.1093/biosci/bix013

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—Evidence from studies of their classroom teachers. *Research in Higher Education*, *34*(2), 151–211.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93–143). Springer. https://doi.org/10.1007/1-4020-5742-3_5

Feldon, D. F., Peugh, J., Maher, M. A., Roksa, J., & Tofel-Grehl, C. (2017). Time-to-credit gender inequities of first-year PhD students in the biological sciences. *CBE Life Sciences Education*, *16*(1), Article ar4. https://doi.org/10.1187/cbe.16-08-0237

Ferber, M. A., & Teiman, M. (1980). Are women economists at a disadvantage in publishing journal articles? *Eastern Economic Journal*, *6*(3/4), 189–193.

Flaherty, K. (2018). *The leaky pipeline for postdocs: A study of the time between receiving a PhD and securing a faculty job for male and female astronomers*. arXiv. https://doi.org/10.48550/arXiv.1810.01511

Forscher, P. S., Cox, W. T. L., Brauer, M., & Devine, P. G. (2019). Little race or gender bias in an experiment of initial review of NIH R01 grant proposals. *Nature Human Behaviour*, *3*(3), 257–264. https://doi.org/10.1038/s41562-018-0517-y

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Walman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), Article eaao0185. https://doi.org/10.1126/science.aao0185

Foschi, M. (1996). Double standards in the evaluation of men and women. *Social Psychology Quarterly*, *59*(3), 237–254. https://doi.org/10.2307/2787021

Foschi, M., Lai, L., & Sigerson, K. (1994). Gender and double standards in the assessments of job candidates. *Social Psychology Quarterly*, *57*, 326–339.

Fowler, C. (2019). Faculty Compensation Survey data and 2018-19 report. *Academe*, *105*(2). https://www.aaup.org/article/faculty-compensation-survey-data-and-2018-19-report#.ZBnJbnbMJOS

Fox, M. F. (2005). Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science*, *35*(1), 131–150. https://doi.org/10.1177/0306312705046630

French, J. C., Zolin, S. J., Lampert, E., Aiello, A., Bencsath, K., Ritter, K., Valente, M., & Probhu, A. S. (2019). Gender and letters of recommendation: A linguistic comparison of the impact of gender on general surgery residency applicants. *Journal of Surgical Education*, *76*, 899–905.

Funk, P., Iriberri, N., & Savio, G. (2019). *Does scarcity of female instructors create demand for diversity among students? Evidence from an M-Turk experiment* (CEPR Discussion Paper No. DP14190). SSRN. https://ssrn.com/abstract=3504620

Gelman, A. (2019, May 11). Did blind orchestra auditions really benefit women? *Statistical Modeling, Causal Inference, and Social Science*. https://statmodeling.stat.columbia.edu/2019/05/11/did-blind-orchestra-auditions-really-benefit-women/

Ginther, D., & Kahn, S. (2014). Academic women's careers in the social sciences. In A. Lanteri & J. Vromen (Eds.), *The economics of economists: Institutional setting, individual incentives, and future prospects* (pp. 285–315). Cambridge University Press. https://doi.org/10.1017/CBO9781139059145.015

Ginther, D. K., & Kahn, S. (2009). Does science promote women? Evidence from academia 1973-2001. In R. B. Freeman & D. L. Goroff (Eds.), *Science and engineering careers in the United States: An analysis of markets and employment* (pp. 163–194). University of Chicago Press.

Ginther, D. K., & Kahn, S. (2013). Education and academic career outcomes for women of color in science and engineering. In K. Matchett (Ed.), *Seeking solutions: Maximizing American talent by advancing women of color in academia* (pp. 71–92). National Academies Press. https://nap.nationalacademies.org/read/18556/chapter/11

Ginther, D. K., & Kahn, S. (2021). Women in academic economics: Have we made progress? *AEA Paper and Proceedings*, *111*, 138–142. https://doi.org/10.1257/pandp.20211027

Ginther, D. K., Kahn, S., & Schaffer, W. T. (2016). Gender, race/ethnicity, and National Institutes of Health R01 research awards: Is there evidence of a double bind for women of color? *Academic Medicine*, *91*(8), 1098–1107.

Glass, C., & Minnotte, K. L. (2010). Recruiting and hiring women in STEM fields. *Journal of Diversity in Higher Education*, *3*(4), 218–229.

Golbeck, A. (2016). Are women underrepresented in the American statistics profession? *Significance*, *13*(2), 8–9. https://doi.org/10.1111/j.1740-9713.2016.00892.x

Goldin, C. (2014). A grand gender convergence: Its last chapter. *The American Economic Review*, *104*(4), 1091–1119.

Goulden, M., Frasch, K., & Mason, M. A. (2009). *Staying competitive: Patching America's leaky pipeline in the sciences*. Center for American Progress. http://cdn.americanprogress.org/wp-content/uploads/issues/2009/11/pdf/women_and_sciences.pdf

Grant, J., Burden, S., & Breen, G. (1997). No evidence of sexism in peer review. *Nature*, *390*(6659), Article 438. https://doi.org/10.1038/37213

Graves, A. L., Hoshino-Browne, E., & Lui, K. P. H. (2017). Swimming against the tide: Gender bias in the physics classroom. *Journal of Women and Minorities in Science and Engineering*, *23*(1), 15–36.

Grogan, K. E. (2019). How the entire scientific community can confront gender bias in the workplace. *Nature Ecology & Evolution*, *3*, 3–6.

Gruber, J., Mendle, J., Lindquist, K. A., Schmader, T., Clark, L. A., Bliss-Moreau, E., Akinola, M., Atlas, L., Barch, D. M., Barrett, L. F., Borelli, J. L., Brannon, T. N., Bunge, S. A., Campos, B., Cantlon, J., Carter, R., Carter-Sowell, A. R., Chen, S., Craske, M. G., . . . Williams, L. A. (2021). The future of women in psychological science. *Perspectives in Psychological Science*, *16*(3), 483–516. https://doi.org/10.1177/1745691620952789

Hamermesh, D. S., & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*, 369–376.

Hammerschmidt, K., Reinhardt, K., & Rolff, J. (2008). Does double-blind review favor female authors? *Frontiers in Ecology and the Environment*, *6*(7), Article 354. https://doi.org/10.1890/1540-9295(2008)6[354a:DDRFFA]2.0.CO;2

Handley, I. M., Brown, E. R., Moss-Racusin, C. A., & Smith, J. L. (2015). Quality of evidence revealing subtle gender biases in science is in the eye of the beholder. *Proceedings of the National Academy of Sciences, USA*, *12*(43), 13201–13206.

Hansson, O. (2009, March). Wold and Wennerås solve 2 unknown variables with 1 equation and receive honorary degree from Chalmers. *Athletic Design*. http://www.athleticdesign.se/otherstuff/sexism_in_peer_review_english.html

Hargens, L. L., & Long, J. S. (2002). Demographic inertia and women's representation among faculty in higher education. *The Journal of Higher Education*, *73*(4), 494–517.

Heath-Stout, L. E. (2020). Guest editorial introduction: Gender, equity, and the peer review process at the Journal of Field Archaeology. *Journal of Field Archaeology*, *45*, 135–139.

Hechtman, L. A., Moore, N. P., Schulkey, C. E., Miklos, A. C., Calcagno, A. M., Aragon, R., & Greenberg, J. H. (2018). NIH funding longevity by gender. *Proceedings of the National Academy of Sciences, USA*, *115*(31), 7943–7948. https://doi.org/10.1073/pnas.1800615115

Heffernan, T. (2022). Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education*, *47*(1), 144–154. https://doi.org/10.1080/02602938.2021.1888075

Heilman, M. E. (1983). Sex bias in work settings: The lack of fit model. *Research in Organizational Behavior*, *5*, 269–298.

Hengel, E. (2022). Publishing while female: Are women held to higher standards? Evidence from peer review. *The Economic Journal*, *132*(648), 2951–2991. https://doi.org/10.1093/ej/ueac032

Henningsen, L., Horvath, L. K., & Jonas, K. (2021). Affirmative action policies in academic job advertisements: Do they facilitate or hinder gender discrimination in hiring processes for professorships? *Sex Roles*, *86*, 34–48. https://doi.org/10.1007/s11199-021-01251-4

Hill, C., Corbett, C., & Rose, R. (2010). *Why so few? Women in science, technology, engineering, and mathematics*. American Association of University Women.

Holliday, E., Jagsi, R., Wilson, R. D., Choi, M., Thomas, C., & Fuller, C. (2014). Gender differences in publication productivity, academic position, career duration, and funding among U.S. academic radiation oncology faculty. *Academic Medicine*, *89*, 767–773. https://doi.org/10.1097/ACM.0000000000000229

Holmes, M. A., Asher, P., Farrington, J., Fine, R., Leinen, M. S., & LeBoy, P. (2011). Does gender bias influence awards given by societies? *Eos*, *92*(47), 421–422. https://doi.org/10.1029/2011EO470002

Hosek, S. D., Cox, A. G., Ghosh-Dastidar, B., Kofner, A., Ramphal, N. R., Scott, J., & Berry, S. H. (2005). *Gender differences in major federal external grant programs*. RAND Corp. https://www.rand.org/pubs/technical_reports/TR307.html

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences, USA*, *117*(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117

Jagsi, R., Motomura, A. R., Griffith, K. A., Rangarajan, S., & Ubel, P. A. (2009). Sex differences in attainment of independent funding by career development awardees. *Annals of Internal Medicine*, *151*(11), 804–811.

Johnson, S. K., Hekman, D. R., & Chan, E. T. (2016, April 26). If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review*. https://hbr.org/2016/04/if-theres-only-one-woman-in-your-candidate-pool-theres-statistically-no-chance-shell-be-hired

Jones, J., Gaffney-Rhys, R., & Jones, E. (2014). Handle with care! An exploration of the potential risks associated with the publication and summative usage of student evaluation of teaching (SET) results. *Journal of Further and Higher Education*, *38*(1), 37–56. https://doi.org/10.1080/0309877X.2012.699514

Jussim, L. (2019, June 23). Scientific bias in favor of studies finding gender bias. *Psychology Today*. https://www.psychologytoday.com/us/blog/rabble-rouser/201906/scientific-bias-in-favor-studies-finding-gender-bias

Kaatz, A., Gutierrez, B., & Carnes, M. (2014). Threats to objectivity in peer review: The case of gender. *Trends in Pharmacological Research*, *35*(8), 371–373. https://doi.org/10.1016/j.tips.2014.06.005

Kaatz, A., Magua, M. W., Zimmerman, D. R., & Carnes, M. (2015). A quantitative linguistic analysis of National Institutes of Health R01 application critiques from investigators at one institution. *Academic Medicine*, *90*(1), 69–75.

Kahn, S. (1993). Gender differences in academic career paths of economists. *The American Economic Review*, *83*(2), 52–56. https://www.jstor.org/stable/2117639

Kahn, S., Ceci, S., & Williams, W. (2022a). *Is there gender bias in grant success? An extended meta-analysis*. SSRN. https://ssrn.com/abstract=4182961

Kahn, S., Ceci, S., & Williams, W. (2022b). *Meta-analysis of gender differences in journal acceptance*. SSRN. https://doi.org/10.2139/ssrn.4182965

Kahn, S. B. (1995). Women in the economics profession. *Journal of Economic Perspectives*, *9*(4), 193–206. https://doi.org/10.1257/jep.9.4.193

Kahn-Lang, A., & Lang, K. (2020). The promise and pitfalls of differences-in-differences: Reflections on *16 and Pregnant* and other applications. *Journal of Business & Economic Statistics*, *38*(3), 613–620. https://doi.org/10.1080/07350015.2018.1546591

Kallens, P. C., Hicks, D. J., & Jennings, C. D. (2022). Networks in philosophy: Social networks and employment in academic philosophy. *Metaphilosophy*, *53*(5), 653–684. https://doi.org/10.1111/meta.12591

Kaminski, D., & Geisler, C. (2012). Survival analysis of faculty retention in science and engineering by gender. *Science*, *335*, 864–866.

Kaufmann, E. (2021). *Academic freedom in crisis: Punishment, political discrimination, and self-censorship*. Center for the Study of Partisanship and Ideology, University of London.

Kelchtermans, S., & Veugelers, R. (2013). Top research productivity and its persistence: Gender as a double-edged sword. *The Review of Economics and Statistics*, *95*(1), 273–285. https://doi.org/10.1162/REST_a_00275

Kessel, C., & Nelson, D. J. (2011). Statistical trends in women's participation in science: Commentary on Valla and Ceci (2011). *Perspectives on Psychological Science*, *6*(2), 147–149.

Key, E. M., & Ardoin, P. (2019). *Gender bias in teaching evaluations: What can be done?* [Paper presentation]. Southern Political Science Association, Austin, TX. https://www.dropbox.com/s/9q7aqa4ka0z4swg/Key%20and%20Ardoin_Evaluation%20Bias_July%202019.pdf?dl=0

Kimura, D. (2002, January 10). Preferential hiring of women. *University of British Columbia Reports*. https://www.safs.ca/newsletters/article.php?article=298

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: An experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, *35*(5), 603–625. https://doi.org/10.1177/1075547012472684

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*(1), 128–161.

Kolev, J., Fuentes-Medel, Y., & Murray, F. (2019). *Is blinded review enough? How gendered outcomes arise even under anonymous evaluation* (NBER Working Paper No. 25759). National Bureau of Economic Research. https://doi.org/10.3386/w25759

Krawczyk, M., & Smyk, M. (2016). Author's gender affects rating of academic articles: Evidence from an incentivized, deception-free laboratory experiment. *European Economics Review*, *90*, 326–335.

Kreitzer, R. J., & Sweet-Cushman, J. (2022). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics*, *20*, 73–84.

Kugler, K. G., Reif, J. A. M., Kaschner, T., & Brodbeck, F. C. (2018). Gender differences in the initiation of negotiations: A meta-analysis. *Psychological Bulletin*, *144*, 198–222. https://doi.org/10.1037/bul0000135

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213.

Lavy, V., & Sand, E. (2015). On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases (NBER Working Paper No. 20909). National Bureau of Economic Research. http://www.nber.org/papers/w20909.pdf

Lawrence, P. A. (2006). Men, women, and ghosts in science. *PLOS Biology*, *4*(1), Article 19. https://doi.org/10.1371/journal.pbio.0040019

Lawson, C., Geuna, A., & Finardi, U. (2021). The funding-productivity-gender nexus in science, a multistage analysis. *Research Policy*, *50*(3), Article 104182. https://doi.org/10.1016/j.respol.2020.104182

Ledin, A., Bornmann, L., Gannon, F., & Wallon, G. (2007). A persistent problem: Traditional gender roles hold back female scientists. *EMBO Reports*, *8*(11), 982–987.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the Association for Information Science and Technology*, *64*(1), 2–17.

Lerchenmueller, M. J., & Sorenson, O. (2018). The gender gap in early career transitions in the life sciences. *Research Policy*, *47*(6), 1007–1017.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265.

Ley, T. J., & Hamilton, B. H. (2008). The gender gap in NIH grant applications. *Science*, *322*(5907), 1472–1474.

Li, D., & Koedel, C. (2017). Representation and salary gaps by race-ethnicity and gender at selective public universities. *Educational Researcher*, *46*(7), 343–354.

Li, S., Fant, A., & McCarthy, D. (2017). Gender differences in language of standardized letters narratives for emergency medical residents. *AEM Education and Training*, *1*, 334–339.

Linek, S. B., Gerjets, P., & Scheiter, K. (2009). The speaker/gender effect: Does the speaker's gender matter when presenting auditory text. *Instructional Science*, *38*, 503–521.

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, *54*, 94–106. https://doi.org/10.1016/j.stueduc.2016.12.004

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *Annals of the American Academy of Political and Social Science*, *587*, 69–81.

Long, J. S. (1992). Measures of sex differences in scientific productivity. *Social Forces*, *71*, 159–178.

Lortie, C. J., Aarssen, L. W., Budden, A. E., Koricheva, J. K., Leimu, R., & Tregenza, T. (2007). Publication bias and merit in ecology. *Oikos*, *116*(7), 1247–1253.

Lubienski, S. T., Miller, E. K., & Saclarides, E. S. (2018). Sex differences in doctoral student publication rates. *Educational Researcher*, *47*(1), 76–81. https://doi.org/10.3102/0013189X17738746

Lutter, M., & Schröder, M. (2016). Who becomes a tenured professor, and why? Panel data evidence from German sociology, 1980–2013. *Research Policy*, *45*(5), 999–1013.

Lynn, F. B., Noonan, M. C., Sauder, M., & Andersson, M. A. (2019). A rare case of gender parity in academia. *Social Forces*, *98*(2), 518–547.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*(4), 291–303.

Madera, J. M., Hebl, M., Dial, H., Martin, R., & Valian, V. (2019). Raising doubt in letters of recommendation for academia. *Journal of Business and Psychology*, *34*, 287–303.

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591–1599.

Madison, G., & Fahlman, P. (2021). Sex differences in the number of scientific publications and citations when attaining the rank of professor in Sweden. *Studies in Higher Education*, *46*(12), 2506–2527. https://doi.org/10.1080/03075079.2020.1723533

Magua, W., Zhu, X., Bhattacharya, A., Filut, A., Potvien, A., Leatherberry, R., Lee, Y., Jens, M., Malikireddy, D., Carnes, M., & Kaatz, A. (2017). Are female applicants disadvantaged in National Institutes of Health peer review? Combining algorithmic text mining and qualitative methods to detect evaluative differences in R01 reviewers' critiques. *Journal of Women's Health*, *26*(5), 560–570.

Maliniak, D., Powers, R., & Walter, B. F. (2013). The gender citation gap in international relations. *International Organization*, *67*(4), 889–922.

Marcotte, A. (2014, December 9). Best way for professors to get good student evaluations? Be male. *Slate*. https://slate.com/human-interest/2014/12/gender-bias-in-student-evaluations-professors-of-online-courses-who-present-as-male-get-better-marks.html

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H.-D., & O'Mara, A. (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research*, *79*(3), 1290–1326. https://doi.org/10.3102/0034654309334143

Martinez, E., Botos, J., Dohoney, K. M., Geiman, T. M., Kolla, S. S., Olivera, A., Qiu, Y., Rayasam, G. V., Stavreva, D. A., & Cohen-Fix, O. (2007). Falling off the academic bandwagon: Women are more likely to quit at the postdoc to principal investigator transition. *EMBO Reports*, *8*(11), 977–981. https://doi.org/10.1038/sj.embor.7401110

Mayer, S., & Rathmann, J. (2018). How does research productivity relate to gender? Analyzing gender differences for multiple publication dimensions. *Scientometrics*, *79*, 1663–1693.

McGillivray, B., & De Ranieri, E. (2018). Uptake and outcome of manuscripts in Nature journals by review model and author characteristics. *Research Integrity and Peer Review*, *3*, Article 5. https://doi.org/10.1186/s41073-018-0049-z

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal*, *35*(1), 37–51.

Mehto, L. I. (2021). The gender gap in highly prestigious international research awards, 2001–2020. *Quantitative Science Studies*, *2*(3), 976–989. https://doi.org/10.1162/qss_a_00148

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535–566. https://doi.org/10.1093/jeea/jvx057

Messner, A. H., & Shimahara, E. (2008). Letters of recommendation to an otolaryngology/head and neck surgery residency program. *Laryngoscope*, *118*, 1335–1344.

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology*, *100*(6), 1678–1712. https://doi.org/10.1037/apl0000022

Miller, C. C. (2015, February 6). Is the professor bossy or brilliant? Much depends on gender. *The New York Times*. https://www.nytimes.com/2015/02/07/upshot/is-the-professor-bossy-or-brilliant-much-depends-on-gender.html

Miller, D. I., Nolla, K. M., Eagly, A. H., & Uttal, D. H. (2018). The development of children's gender-science stereotypes: A meta-analysis of 5 decades of U.S. Draw-A-Scientist studies. *Child Development*, *89*(6), 1943–1955. https://doi.org/10.1111/cdev.13039

Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology*, *28*(4), 283–299.

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, *51*(3), 648–652.

Moratti, S. (2020). A woman's place is in the 'home'? Gender-specific hiring patterns in academia in gender-equal Norway. *Journal of Sociology*, *57*(4), 916–934. https://doi.org/10.1177/1440783320960530

Morgan, A. C., Way, S. F., Hoefer, M. D., Larremore, D. B., Galesic, M., & Clauset, A. (2021). The unequal impact of parenthood in academia. *Science Advances*, *7*(9), Article eabd1996. https://doi.org/10.1126/sciadv.abd1996

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, USA*, *109*(41), 16474–16479.

Murray, D., Siler, K., Lariviére, V., Chan, W. M., Collings, A. M., Raymond, J., & Sugimoto, C. R. (2019). *Author-reviewer homophily in peer review*. bioRxiv. https://doi.org/10.1101/400515

Mutz, R., Bornmann, L., & Daniel, H. (2012). Does gender matter in grant peer review? *Zeitschrift für Psychologie*, *220*(2), 121–129. https://doi.org/10.1027/2151-2604/a000103

National Academies of Sciences, Engineering, and Medicine. (2018). *Sexual harassment of women: Climate, culture, and consequences in academic sciences, engineering, and medicine*. The National Academies Press. https://doi.org/10.17226/24994

National Academies of Sciences, Engineering, and Medicine. (2020a). *Promising practices for addressing the underrepresentation of women in science, engineering, and medicine: Opening doors*. The National Academies Press. https://doi.org/10.17226/25585

National Academies of Sciences, Engineering, and Medicine. (2020b). *Promising practices for addressing the underrepresentation of women in science, engineering, and medicine: Opening doors: Highlights*. The National Academies Press. https://nap.nationalacademies.org/resource/25585/ReportHighlights_Promising%20Practices.pdf

National Academy of Sciences. (2007). *Beyond bias and barriers: Fulfilling the potential of women in academic science and engineering*. The National Academies Press. https://doi.org/10.17226/11741

National Institutes of Health. (2020). *NIH data book: R01-equivalent grants: Success rates, by gender and type of application*. https://report.nih.gov/nihdatabook/report/131

National Research Council. (2010). *Gender differences at critical transitions in the careers of science, engineering, and mathematics faculty*. The National Academies Press.

Netherlands Organisation for Scientific Research. (2023a). *Vici 2019*. https://www.nwo.nl/en/researchprogrammes/nwo-talent-programme/projects-vici/2019

Netherlands Organisation for Scientific Research. (2023b). *Vidi 2019*. https://www.nwo.nl/en/onderzoeksprogrammas/nwo-talent-programme/projects-vidi/2019

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101

Odic, D., & Wojcik, E. H. (2020). The publication gender gap in psychology. *American Psychologist*, *75*(1), 92–103. https://doi.org/10.1037/amp0000480

O'Meara, K., Culpepper, D., & Templeton, L. L. (2020). Nudging toward diversity: Applying behavioral design to faculty hiring. *Review of Educational Research*, *90*(3), 311–348. https://doi.org/10.3102/0034654320914742

Østby, G., Strand, H., Nordås, R., & Gleditsch, N. P. (2013). Gender gap or gender bias in peace research? Publication patterns and citation rates for *Journal of Peace Research*, 1983–2008. *International Studies Perspectives*, *14*(4), 493–506. https://doi.org/10.1111/insp.12025

Pallesen, J. S.-M. (2019, May 11). Blind auditions and gender discrimination. JSMP. http://jsmp.dk/posts/2019-05-12-blindauditions/

Pennebaker, J., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC 2015*. University of Texas at Austin. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

Perc, M. (2014). The Matthew effect in empirical data. *Journal of the Royal Society Interface*, *11*(98), Article 20140378. https://doi.org/10.1098/rsif.2014.0378

Perna, L. (2001). Sex differences in faculty salaries: A cohort analysis. *Review of Higher Education*, *24*(3), 283–307.

Pezzoni, M., Mairesse, J., Stephan, P., & Lane, J. (2016). Gender and the publication output of graduate students: A case study. *PLOS ONE*, *11*(1), Article e0145146. https://doi.org/10.1371/journal.pone.0145146

Pohlhaus, J. R., Jiang, H., Wagner, R. M., Schaffer, W. T., & Pinn, V. W. (2011). Sex differences in application, success, and funding rates for NIH extramural programs. *Academic Medicine*, *86*(6), 759–767.

Posselt, J. R. (2016). *Inside graduate admissions: Merit, diversity, and faculty gatekeeping*. Harvard University Press. https://doi.org/10.4159/9780674915640

Price, L., Svensson, I., Borell, J., & Richardson, J. T. (2017). The role of gender in students' ratings of teaching quality in computer science and environmental engineering. *IEEE Transactions on Education*, *60*(4), 281–287.

Reid, I. N. (2014). Gender-correlated systematics in HST proposal selection. *Publications of the Astronomical Society of the Pacific*, *126*(944), Article 923. https://doi.org/10.1086/678964

Reif, J., Kunz, F., Kugler, K., & Brodbeck, F. (2019). Negotiation contexts: How and why they shape women's

and men's decision to negotiate. *Negotiation and Conflict Management Research*, *12*(4), 343–366. https://doi.org/10.34891/6kj3-9093

Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences, USA*, *111*, 4403–4408.

Rissler, L. J., Hale, K. L., Joffe, N. R., & Caruso, N. M. (2020). Gender differences in grant submissions across science and engineering fields at the NSF. *BioScience*, *70*(9), 814–820. https://doi.org/10.1093/biosci/biaa072

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review*, *84*(2), 248–274.

Roberts, S. G., & Verhoef, T. (2016). Double-blind reviewing at EvoLang 11 reveals gender bias. *Journal of Language Evolution*, *1*(2), 163–167. https://doi.org/10.1093/jole/lzw009

Rockey, S. (2014, August 8). Women in biomedical research. *NIH Extramural Nexus*. https://nexus.od.nih.gov/all/2014/08/08/women-in-biomedical-research/

Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: A large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education*, *43*(1), 31–44.

Ross, M. B., Glennon, B. M., Murciano-Goroff, R., Berkes, E. G., Weinberg, B. A., & Lane, J. L. (2022). Women are credited less in science than men. *Nature*, *608*, 135–145. https://doi.org/10.1038/s41586-022-04966-w

Royal Society of Chemistry. (2019). *Is publishing in the chemical sciences gender biased? Driving change in research culture*. https://www.rsc.org/globalassets/04-campaigning-outreach/campaigning/gender-bias/gender-bias-report-final.pdf

Sakai, A. K., & Lane, M. J. (1996). National Science Foundation funding patterns of women and minorities in biology. *BioScience*, *46*, 621–625.

Salinas, P. C., & Bagni, C. (2017). Gender equality from a European perspective: Myth and reality. *Neuron*, *96*(4), 721–729.

Samaniego, C., Lindner, P., Kazmi, M. A., Dirr, B.A., Kong, D. T., Jeff-Eke, E., & Spitzmueller, C. (2023). Higher research productivity = more pay? Gender pay-for-productivity inequity across disciplines. *Scientometrics*, *128*, 1395–1407. https://doi.org/10.1007/s11192-022-04513-4

Sandström, U., & Hällsten, M. (2008). Persistent nepotism in peer-review. *Scientometrics*, *74*(2), 175–189.

Saul, J. (2009, October 7). Women in philosophy: What's getting left out. *Feminist Philosophers*. https://feministphilosophers.wordpress.com/2009/10/07/women-in-philosophy-whats-getting-left-out/

Sax, L. J., Hagedorn, L. S., Arredondo, M., & Dicrisi, F. A. (2002). Faculty research productivity: Exploring the role of gender and family-related factors. *Research in Higher Education*, *43*, 423–446.

Schaerer, M., du Plessis, C., Nguyen, M. H. B., van Aert, R. C. M., Tiokhin, L., Lakens, D., Clemente, E. G., Pfeiffer, T., Dreber, A., Johannesson, M., Clark, C. J., Gender Audits Forecasting Collaboration, & Uhlmann, E. L. (2022). *On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions* [Manuscript submitted for publication]. Columbia University.

Schaller, M. D. (2023). Research productivity and training support for doctoral students in the biological and biomedical sciences. *BioAdvances*, *5*(3), 131–148. https://doi.org/10.1096/fba.2022-00117

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, *57*(7–8), 509–561.

Schmidt, B. (2020). *Gendered language in teacher reviews*. http://benschmidt.org/profGender/#

Schröder, M., Lutter, M., & Habicht, I. M. (2021). Publishing, signaling, social capital, and gender: Determinants of becoming a tenured professor in German political science. *PLOS ONE*, *16*(1), Article e0243514. https://doi.org/10.1371/journal.pone.0243514

Schroeder, J., Dugdale, H. L., Radersma, R., Hinsch, M., Buehler, D. M., Saul, J., Porter, L., Liker, A., Decauwer, I., Johnson, P. J., Santure, A. W., Griffin, A. S., Bolund, E., Ross, L., Webb, T. J., Feulner, P. G. D., Winney, I., Szulkin, M., Komdeur, J., . . . Horrocks, N. P. C. (2014). Fewer invited talks by women in evolutionary biology symposia. *Journal of Evolutionary Biology*, *28*, 2063–2069.

Schubert, F., & Engelage, S. (2011). Wie undicht ist die pipeline? Wissenschaftskarrieren von promovierten Frauen. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *63*, 431–457. https://doi.org/10.1007/s11577-011-0144-3

Seldin, P. (1993, September 8). The use and abuse of student ratings of professors. *The Chronicle of Higher Education*, *39*(46), Article A40. https://www.chronicle.com/article/the-use-and-abuse-of-student-ratings/

Settles, I. H., Cortina, L. M., Malley, J., & Stewart, A. J. (2006). The climate for women in academic science: The good, the bad, and the changeable. *Psychology of Women Quarterly*, *30*, 47–58.

Severin, A., Martins, J., Heyard, R., Delavy, F., Jorstad, A., & Egger, M. (2020). Gender and other potential biases in peer review. *BMJ Open*, *10*, Article e035058. https://doi.org/10.1136/bmjopen-2019-035058

Sheehan, E. P., McDevitt, T. M., & Ross, H. C. (1998). Looking for a job as a psychology professor? Factors affecting applicant success. *Teaching of Psychology*, *25*(1), 8–11. https://doi.org/10.1207/s15328023top2501_3

Sheltzer, J. M., & Smith, J. C. (2014). Elite male faculty in the life sciences employ fewer women. *Proceedings of the National Academy of Sciences, USA*, *111*(28), 10107–10112.

Shen, H. (2013). Inequality quantified: Mind the gender gap. *Nature*, *495*(7439), 22–24. https://doi.org/10.1038/495022a

Skibba, R. (2019). Women in physics. *Nature Reviews Physics*, *1*, 298–300.

Smyth, F. L., & Nosek, B. A. (2015). On the gender–science stereotypes held by scientists: Explicit accord with gender-ratios, implicit accord with scientific identity. *Frontiers in Psychology*, *6*, Article 415. https://doi.org/10.3389/fpsyg.2015.00415

Snyder, T. D. (Ed.). (1993). *120 years of American education: A statistical portrait*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/pubs93/93442.pdf

Snyder, T. D., de Brey, C., & Dillow, S. A. (2019). *Digest of education statistics 2018*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/pubs2020/2020009.pdf

Solga, H., Rusconi, A., & Netz, N. (2023). Professors' gender biases in assessing applicants for professorships. *European Sociological Review*. Advance online publication. https://doi.org/10.1093/esr/jcad007

Sommers, C. H. (2019, October 20). Blind spots in the 'blind audition' study. *The Wall Street Journal*. https://www.wsj.com/articles/blind-spots-in-the-blind-audition-study-11571599303

Squazzoni, F., Bravo, G., Farjam, M., Marusic, A., Mehmani, B., Willis, M., Birukou, A., Dondio, P., & Grimaldo, F. (2021). Peer review and gender bias: A study on 145 scholarly journals. *Science Advances*, *7*(2), Article eabd0299. https://doi.org/10.1126/sciadv.abd0299

Stankovic, J., & Aspray, W. (2003). *Recruitment and retention of faculty in computer science and engineering*. Computer Research Association. http://archive.cra.org/reports/r&rfaculty.pdf

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*(7–8), 509–528. https://doi.org/10.1023/A:1018839203698

Stewart, P., Ornstein, M., & Drakich, J. (2009). Gender and promotion at Canadian universities. *Canadian Review of Sociology/Revue Canadienne de Sociologie*, *46*(1), 59–85.

Stone, D. L., & Rosopa, P. J. (2017). The advantages and limitations of using meta-analysis in human resource management research. *Human Resource Management Review*, *27*, 1–7. https://doi.org/10.1016/j.hrmr.2016.09.001

Storage, D., Horne, Z., Cimpian, A., & Leslie, S.-J. (2016). The frequency of "brilliant" and "genius" in teaching evaluations predicts the representation of women and African Americans across fields. *PLOS ONE*, *11*(3), Article e0150194. https://doi.org/10.1371/journal.pone.0150194

Strolger, L., & Natarajan, P. (2019). Doling out *Hubble* time with dual-anonymous evaluation. *Physics Today*. https://doi.org/10.1063/PT.6.3.20190301a

Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and thing interests explain gender disparities across STEM fields. *Frontiers in Psychology*, *6*, Article 189. https://doi.org/10.3389/fpsyg.2015.00189

Su, R., Rounds, J., & Armstrong, P. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*, 859–884.

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

Symonds, M. R. E., Gemmell, N. J., Braisher, T. L., Gorringe, K. L., & Elgar, M. A. (2006). Gender differences in publication output: Towards an unbiased metric of research performance. *PLOS ONE*, *1*(1), Article e127. https://doi.org/10.1371/journal.pone.0000127

Tamblyn, R., Girard, N., Qian, C. J., & Hanley, J. (2018). Assessment of potential bias in research grant peer review in Canada. *Canadian Medical Association Journal*, *190*(16), E489–E499.

Teich, E. G., Kim, J. Z., Lynn, C. W., Simon, S. C., Klishin, A. A., Szymula, K. P., Srivastava, P., Bassett, L. C., Zurn, P., Dworkin, J. D., & Bassett, D. S. (2022). Citation inequity and gendered citation practices in contemporary physics. *Nature Physics*, *18*, 1161–1170. https://doi.org/10.1038/s41567-022-01770-1

Tekles, A., Auspurg, K., & Bornmann, L. (2022). Same-gender citations do not indicate a substantial gender homophily bias. *PLOS ONE*, *17*(9), Article e0274810. https://doi.org/10.1371/journal.pone.0274810

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, *77*, Article 101981. https://doi.org/10.1016/j.econedurev.2020.101981

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017a). Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences, USA*, *114*(48), 12708–12713. https://doi.org/10.1073/pnas.1707323114

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017b). *Single versus double blind reviewing at WSDM 2017*. arXiv. https://arxiv.org/pdf/1702.00502v1.pdf

Trevino, L. J., Balkin, D. B., & Gomez-Mejia, L. R. (2017). How doing gender leads to gender imbalances in the higher ranks of colleges and business. *Academy of Management, Learning and Education*, *16*, 439–443.

Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, *14*(2), 191–220.

Tung, A. K. (2006). Impact of double-blind reviewing on SIGMOD publication: A more detail analysis. *ACM SIGMOD Record*, *35*(3), 6–7.

The University of Western Ontario, Office of the Provost and Vice-President (Academic). (2000). Figure 11: Western: % women in recruitment activities leading to new annual tenured, probationary, and limited-term faculty appointments (excludes clinicians). In *Full-time faculty distribution, appointments, and recruitment – by gender* (p. 12). https://www.uwo.ca/western/facgend99/facgend.pdf

U.S. Government Accountability Office. (2015). *Women in STEM research: Better data and information sharing could improve oversight of federal grant-making and Title IX Compliance*. Government Accountability Office. www.gao.gov/products/GAO-16-14

Urquhart-Cronish, M., & Otto, S. P. (2019). Gender and language use in scientific grant writing. *Facets*, *4*(1), 442–458. https://doi.org/10.1139/facets-2018-0039

Urry, M. (2015). Science and gender: Scientists must work harder on equality. *Nature*, *528*, 471–473.

Uttl, B. (2021). Lessons learned from research on student evaluation of teaching in higher education. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in*

schools (pp. 237–256). Springer. https://doi.org/10.1007/978-3-030-75150-0_15

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*, *5*, Article e3299. https://doi.org/10.7717/peerj.3299

Uttl, B., & Violo, V. (2021). Gender bias in student evaluation of teaching or a mirage? *ScienceOpen Research*. https://doi.org/10.14293/S2199-1006.1.SOR.2021.0003.v1

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Valkonen, L., & Brooks, J. (2011). Gender balance in Cortex acceptance rates. *Cortex*, *47*(7), 763–770. https://doi.org/10.1016/j.cortex.2011.04.004

van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2012). Gender differences in scientific productivity: A persisting phenomenon? *Scientometrics*, *93*(3), 857–868.

van den Besselaar, P., & Sandström, U. (2016). Gender differences in research performance and its impact on careers: A longitudinal case study. *Scientometrics*, *106*(1), 143–162.

van den Besselaar, P., & Sandström, U. (2017). Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PLOS ONE*, *12*(8), Article e0183301. https://doi.org/10.1371/journal.pone.0183301

van der Lee, R., & Ellemers, N. (2015a). Gender contributes to personal research funding success in The Netherlands. *Proceedings of the National Academy of Sciences, USA*, *112*(40), 12349–12345. https://doi.org/10.1073/pnas.1510159112

van der Lee, R., & Ellemers, N. (2015b). Reply to Albers: Acceptance of empirical evidence for gender disparities in Dutch research funding. *Proceedings of the National Academy of Sciences, USA*, *112*(50), Article E6830. https://doi.org/10.1073/pnas.1521336112

Van Miegroet, H., & Glass, C. (2020). Recognition through awards: A source of gender inequality in science? *International Journal of Gender, Science and Technology*, *12*(2).

von Hippel, W., & Buss, D. M. (2017). Do ideologically driven scientific agendas impede the understanding and acceptance of evolutionary principles in social psychology? In J. T. Crawford & L. Jussim (Eds.), *Frontiers of Social Psychology Series: The politics of social psychology* (pp. 7–25). Routledge.

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review*, *54*, 79–94. https://doi.org/10.1016/j.econedurev.2016.06.004

Waisbren, S. E., Bowles, H., Hasan, T., Zou, K. H., Emans, S. J., Goldberg, C., Gould, S., Levine, D., Lieberman, E., Loeken, M., Longtine, J., Nadelson, C., Patenaude, A. F., Quinn, D., Randolph, A. G., Solet, J. M., Ullrich, N., Walensky, R., Weitzman, P., & Christou, H. (2008). Gender differences in research grant applications and funding outcomes for medical school faculty. *Journal of Women's Health*, *7*(2), 207–214. https://doi.org/10.1089/jwh.2007.0412

Walker, R., Barros, B., Conejo, R., Neumann, K., & Telefont, M. (2015). Personal attributes of authors and reviewers, social bias and the outcomes of peer review: A case study. *F1000Research*, *4*, Article 21. https://doi.org/10.12688/f1000research.6012.2

Wallisch, P., & Cachia, J. Y. A. (2019). *Determinants of perceived teaching quality: The role of divergent interpretations of expectations*. PsyArXiv. https://doi.org/10.31234/osf.io/dsvgq

Warner, E. T., Carapinha, R., Weber, G. M., Hill, E. V., & Reede, J. Y. (2017). Gender differences in receipt of National Institutes of Health R01 grants among junior faculty at an academic medical center: The role of connectivity, rank, and research productivity. *Journal of Women's Health*, *26*(10), 1086–1093.

Way, S. F., Larremore, D. B., & Clauset, A. (2016). *Gender, productivity, and prestige in computer science faculty hiring networks*. ArXiv. https://doi.org/10.48550/arXiv.1602.00795

Webb, T. J., O'Hara, B., & Freckleton, R. P. (2008). Does double-blind review benefit female authors? *Trends in Ecology & Evolution*, *23*(7), 351–353.

Webber, K. L., & Canché, M. G. (2018). Is there a gendered path to tenure? A multi-state approach to examine the academic trajectories of U.S. doctoral recipients in the sciences. *Research in Higher Education*, *59*(7), 897–932.

Weisshaar, K. (2017). Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces*, *96*, 529–560. https://doi.org/10.1093/sf/sox052

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature*, *387*, 341–343.

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLOS ONE*, *8*(7), Article e66212. https://doi.org/10.1371/journal.pone.0066212

Whittaker, R. J. (2008). Journal review and gender equality: A critical comment on Budden et al. *Trends in Ecology & Evolution*, *23*(9), 478–479. https://doi.org/10.1016/j.tree.2008.06.003

Williams, W. M., & Ceci, S. J. (2012). When scientists choose motherhood. *American Scientist*, *100*, 138–145.

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences, USA*, *112*(17), 5360–5365.

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2018). *Female grant applicants are equally successful when peer reviewers assess the science, but not when they assess the scientist*. bioRxiv. https://doi.org/10.1101/232868

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet*, *393*(10171), 531–540. https://doi.org/10.1016/S0140-6736(18)32611-4

Witze, A. (2020). Three extraordinary women run the gauntlet of science—a documentary. *Nature*, *583*, 25–26. https://doi.org/10.1038/d41586-020-01912-6

Wold, A., & Chrapkowska, C. (2004). *Förbjuden frukt på kunskapens träd* [Forbidden fruit on the tree of knowledge]. Atlantis.

Wolfinger, N. H., Mason, M. A., & Goulden, M. (2008). Problems in the pipeline: Gender, marriage, and fertility in the ivory tower. *The Journal of Higher Education*, *79*(4), 388–405. https://doi.org/10.1080/00221546.2008.11772108

Wong, J. S., & Bouchard, J. (2021). Are students gender-neutral in their assessment of online teaching staff? *Assessment & Evaluation in Higher Education*, *46*(5), 719–739. https://doi.org/10.1080/02602938.2020.1805410

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education*, *37*(6), 683–699.

Xie, Y., & Shauman, K. A. (1998). Sex differences in research productivity: New evidence about an old puzzle. *American Sociological Review*, *63*, 847–870.

Xie, Y., & Shauman, K. A. (2003). *Women in science: Career processes and outcomes*. Harvard University Press.

Xu, Y. (2015). Focusing on women in STEM: A longitudinal examination of gender-based earning gap of college graduates. *The Journal of Higher Education*, *86*(4), 489–523. https://doi.org/10.1353/jhe.2015.0020

Zigerell, L. J. (2018, August 28). *Is there a citation bias against research findings that undercut liberal assumptions or policy preferences?* https://www.ljzigerell.com/?p=4840

Zigerell, L. J. (2021, January 31). *Ideology and vote choice of election experts in the U.S. 2016 Perceptions of Electoral Integrity study*. https://www.ljzigerell.com/?cat=49