SOCIAL SCIENCES

Peer review and gender bias: A study on 145 scholarly journals

Flaminio Squazzoni[1]*, Giangiacomo Bravo[2], Mike Farjam[3], Ana Marusic[4], Bahar Mehmani[5], Michael Willis[6], Aliaksandr Birukou[7], Pierpaolo Dondio[8], Francisco Grimaldo[9]

Scholarly journals are often blamed for a gender gap in publication rates, but it is unclear whether peer review and editorial processes contribute to it. This article examines gender bias in peer review with data for 145 journals in various fields of research, including about 1.7 million authors and 740,000 referees. We reconstructed three possible sources of bias, i.e., the editorial selection of referees, referee recommendations, and editorial decisions, and examined all their possible relationships. Results showed that manuscripts written by women as solo authors or coauthored by women were treated even more favorably by referees and editors. Although there were some differences between fields of research, our findings suggest that peer review and editorial processes do not penalize manuscripts by women. However, increasing gender diversity in editorial teams and referee pools could help journals inform potential authors about their attention to these factors and so stimulate participation by women.

INTRODUCTION

The academic publishing system shows a systematic underrepresentation of women as authors, referees, and editors (*1*). This underrepresentation is persistent (*2*, *3*) and well documented in various fields of research (*4*–*6*). While some previous studies have found no substantial productivity gap in specific fields, in more recent cohorts of academics or when using less biased output measures (*7*, *8*), a recent study of 1.5 million academics suggested that the relative increase of participation of women in science, technology, engineering, and mathematics (STEM) fields over the past 60 years has not reduced the gap in women's academic productivity and impact (*9*). Even in fields such as the humanities, psychology, and the social sciences, where the gender composition of the community has been more favorable to women for decades, men still publish more manuscripts and in more prestigious journals (*10*, *11*). In the current hypercompetitive academic environment, such a publication gap could explain why women often have a higher probability of dropout from academia, fewer grants, lower salaries, and less prestigious careers (*1*, *12*).

In this context, scholarly journals are often blamed for this gender gap (*13*, *14*). However, whether peer review and journal editorial processes are the root cause of these gender penalties is disputed (*15*). On the one hand, recent reports from journals in specific fields, especially in political science, suggest that editorial processes do not discriminate against women (*16*–*19*). For instance, a recent study of four leading journals in economics also found negligible effects of gender on the assessment of manuscripts (*20*). On the other hand, recent research in other fields, such as ecology, found that manuscripts submitted by women as first authors received slightly worse peer review scores and were more likely to be rejected after peer review (*21*). While the publication gap between men and women is generally explained by persistent differences in submission rates by women in almost all fields of research, it is unclear whether peer review and editorial processes contribute to it.

Furthermore, the fact that women are systematically less involved in peer review and are rarely appointed to prestigious editorial positions (*13*, *14*, *22*) could influence women's perceptions of their adequacy and potential success as authors. For instance, recent research suggests that women would submit fewer manuscripts, of comparably higher quality than those written by men, because they anticipate possible editorial bias and invest more in their manuscripts (*23*). A recent survey of a sample of 2440 American Political Science Association members revealed that women prefer not to target certain journals as they perceive that they will have lower chances than men with similar expertise (*24*).

Unfortunately, establishing whether peer review and editorial processes have any direct or indirect effect on the lower rate of publications by women is difficult (*21*, *25*). It is likely that previous research did not achieve a consensus in findings because data were either case specific or could not capture all the internal steps at journals that might reveal potential bias. The fact that research has never been performed at a scale sufficient to provide insights in different fields of research and journal contexts has made comparison difficult and has not helped to understand whether specific models of peer review, e.g., single versus double blind, could trigger gender bias (*26*).

Our study aims to fill this gap by providing the first in-depth analysis of peer review and editorial processes in a large sample of scholarly journals in different fields of research that considers editorial processes as a set of interlinked decisions. We concentrated on three possible sources of bias, i.e., the editorial selection of referees, referee recommendations, and editorial decisions, and examined all their possible relationships while controlling for important confounding factors such as journals' field of research, impact factor, and single- versus double-blind peer review. Because of an agreement on data sharing with some of the largest scholarly publishers (*27*), we collected complete and fully comparable temporal data on 145 scholarly journals, including almost 350,000 submissions by about 1.7 million authors and more than 760,000 reviews performed by about 740,000 referees (see Materials and Methods). To the best

[1]Department of Social and Political Sciences, University of Milan, Milan, Italy. [2]Department of Social Studies and Centre for Data Intensive Sciences and Applications, Linnaeus University, Växjö, Sweden. [3]Department of Computer Science and Media Technology and Centre for Data Intensive Sciences and Applications, Linnaeus University, Växjö, Sweden. [4]University of Split School of Medicine, Split, Croatia. [5]STM Journals, Elsevier, Amsterdam, Netherlands. [6]John Wiley & Sons, Oxford, UK. [7]Springer Nature, Heidelberg, Germany. [8]School of Computer Science, Technological University Dublin, Dublin, Ireland. [9]Department of Computer Science, University of Valencia, Burjassot, Spain.
*Corresponding author. Email: flaminio.squazzoni@unimi.it

of our knowledge, this is the first study that includes data on manuscripts and reviewer scores across journals from different publishers and fields of research of sufficient depth to assess whether peer review and editorial process contribute to the gender gap in publications.

RESULTS

Table 1 shows the distribution of journals by fields of research in our sample, the proportion of women among authors, and other summary statistics. Our data confirmed previous research on gender disparities in manuscript submissions and peer reviewing (*13*, *14*, *18*, *28*), with 75% of men among submission authors and 79% of men among referees. As expected, we found differences between journals from different research fields, with the greater gender gap in the rate of women among authors and referees in physics (only 19% women as authors). In addition, women are less involved in peer review compared to their authorship rate in all domains except for social sciences (38% women as authors and referees; Table 1). While this could reflect the different rate of adoption of diversity and inclusion policies in some journals, it is more probable that these distortions simply reflect differences in the gender composition of the potential pool of authors and referees, which is impossible to estimate.

Figure 1 shows an overview of the distribution of the final editorial decisions on manuscripts by gender of the first and last author and field of research. This picture suggests a certain degree of diversity among fields, e.g., manuscripts by women would be accepted more frequently in biomedicine, health science, and social science journals, and less frequently in life science journals. However, these descriptive statistics do not allow us to consider the potential effect of important covariates such as the journal's impact factor, the number of coauthors, and the review scores, which would be essential in untangling potential sources of bias during the editorial and peer review process. Note that data on desk rejections were not consistently available, and so, we concentrated on manuscripts that were not desk-rejected by editors.

To examine these processes more systematically, we performed robust statistical analysis within a Bayesian framework and estimated different models on the dataset (see Materials and Methods). We first looked at the editorial process by considering each of the following steps separately: (i) the editorial selection of referees, (ii) the referee recommendations, and (iii) the editorial decision on the manuscript. All these steps included specific actions performed by either referees or editors that could reveal a bias. Following previous research and based on data availability, we considered both the position of women in the author list (i.e., whether they were first or last authors) and the proportion of women among the authors as main predictors (*10*, *13*, *21*) while controlling for the proportion of women among the referees, the impact factor of the journal, the number of authors in each manuscript, and the type of peer review adopted by the journal (*29*, *30*).

Given that the effect of many of these variables, and crucially of the gender of first and last authors, is likely to be different in each field of research, we estimated separated models for each field. This allowed us to consider field specificities, including the journal prestige and potential diversity of evaluation standards, through in-depth data that have never been available before in this type of research (*15*). We then built a Bayesian-learning network model (*31*) to estimate the effect of complex interactions more systematically and understand the extent and persistence of gender bias across all steps of the editorial process (see Materials and Methods).

Regarding the editorial selection of referees (step i), we found that in all fields of research, manuscripts with a higher proportion of women among the authors were more usually reviewed by women referees (see table S1). This is consistent with previous research (*13*) and was confirmed after controlling for the number of authors in the manuscript, the journal's impact factor, and the type of peer review model (single versus double blind). Whether such author-referee gender matching is due to any intentional preference or deliberate practice of journal editors or simply reflects an unequal distribution of men and women in expertise and fields of research is beyond the scope of this study. Our findings here simply indicate that manuscripts by women were not differently treated because of being usually reviewed by men.

Furthermore (step ii), we found that manuscripts by women received systematically more positive reviews in biomedicine and health sciences, as well as in social sciences, whereas they were less positively treated in life sciences (weak statistical effect) and physical sciences journals (strong statistical effect). Women tended to provide more positive recommendations than men in all fields but physical sciences. This effect was consistent after controlling for all other variables and can therefore not be explained by the gender matching between referees and authors or other factors (see table S2). The fact that our model could only explain a small fraction of the outcome variance (between 4 and 11%, depending on the field of research), though many model coefficients were significant, suggests that other manuscript characteristics that we could not measure,

Table 1. Number of journals and frequency distribution of selected sample characteristics by field of research.

	Biomedicine and health	Life sciences	Physical sciences	Social sciences and humanities
Number of journals	55	24	50	16
Mean impact factor (SD)	2.99 (1.49)	3.14 (1.60)	3.04 (1.32)	2.18 (1.07)
Number of submissions	113,421	31,331	184,315	19,051
Percentage first-round rejections	45.8	35.2	41.2	50.0
Percentage final rejections	58.8	48.1	48.5	62.3
Percentage women authors	31.5	27.7	19.1	38.0
Percentage women referees	24.6	21.0	16.3	38.1

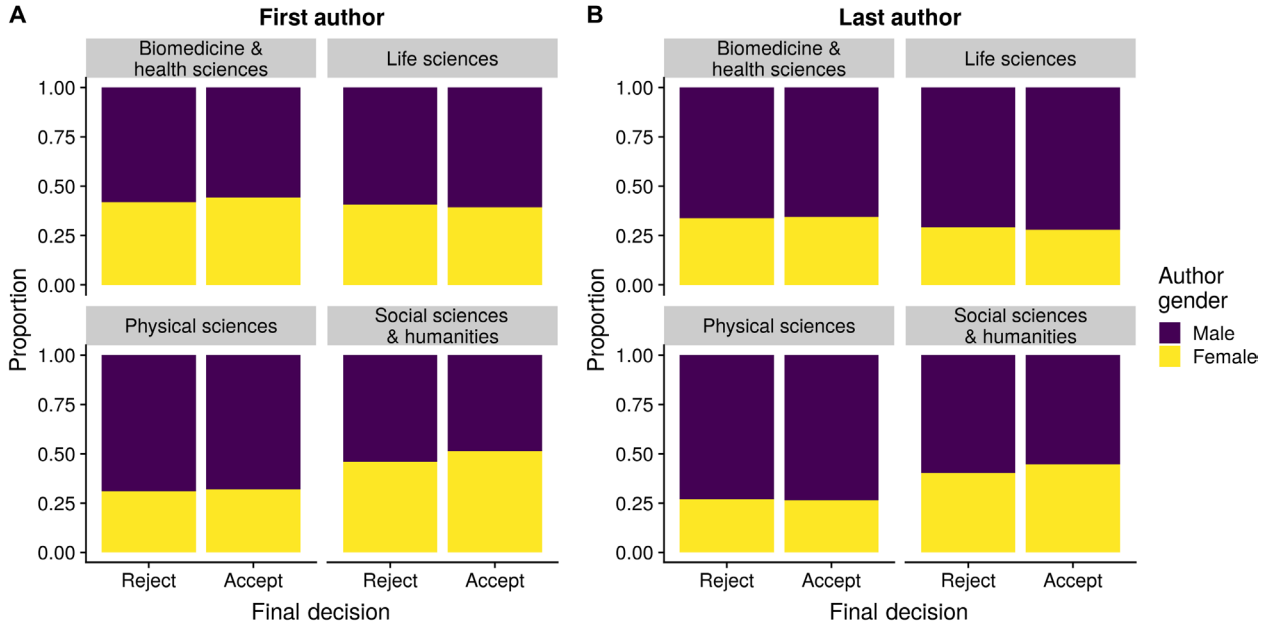

Fig. 1. Distribution of final editorial decisions of manuscripts that were sent out for review by the gender of the first and last author.

such as its quality and content, had the strongest effect on referee recommendations. This effect was independent of any editorial matching or referee selection options.

To check whether our results were robust in taking into consideration alternative specifications of our gender variable, we estimated two further models that considered (i) whether a woman was first or last author of a manuscript (see table S3) and (ii) whether effects were different for our five mutually exclusive groups of authors: a man sole author, a woman sole author, all-men teams, all-women teams, and co-ed teams of authors (see table S4) (*10*). In general, our results show that the author gender did not have a consistent effect, although we found the emergence of complex patterns of interaction when the field of research of journals and the specific composition of author groups were taken into account (for a more systematic analysis of these complex interactions, see the Bayesian-learning network below).

Regarding the final editorial decisions (step iii), we found that manuscripts with a higher proportion of women among authors were accepted more frequently in biomedical, health sciences, and physical sciences journals (strong statistical effect), whereas no evidence of any effect of the gender variable was found in life sciences and social sciences journals. Note that in case of biomedical and physical sciences journals, the positive effect was robust across variation of contexts and controlling for the referee recommendations and the journal's field of research (Table 2). Furthermore, considering the review scores (for details on referee recommendations, see Materials and Methods), our models were able to explain over 80% of the outcome variance.

Alternative specifications of the gender variable did not lead to any systematic difference in the gender effects mentioned above, although resulting in less clear-cut results than in previous models (Table 3). When we considered the gender of the first author, we found that manuscripts by women were more favorably treated in physical sciences journals (strong statistical effect) and less in life sciences journals (weak statistical effect). Being the last author had no significant effect on acceptance, except for a weak negative effect in case of biomedical and health sciences journals. We did not find any systematic bias against manuscripts submitted by women across journals and disciplines when considering the four author groups mentioned above (see table S5).

Last, to consider the whole editorial process in which indirect opportunities for bias may exist and assuming that complex interactions among variables could affect editorial decisions, we estimated a Bayesian-learning network including all the previous steps of the analysis. After learning coefficients and conditional probabilities through maximum likelihood estimation, our model was able to predict with 82% accuracy whether or not a manuscript would ultimately be accepted by the editor (see Materials and Methods). Figure 2 shows that after controlling for all direct and indirect effects of all variables, the effect of authors' gender on referee recommendations depended on the field of research. While manuscripts with a higher proportion of women among authors received slightly more positive recommendations in journals from social sciences and biomedical and health sciences, referee recommendations were slightly more negative for manuscripts submitted to life and physical sciences journals. However, even when comparing the extreme cases where manuscripts were authored exclusively by women or men, our model predicted a change in review scores by less than 4%, showing that these effects were minimal.

Note that while the directionality of paths is necessary to estimate path coefficients in Bayesian networks, the direction of arrows does not necessarily imply causation (*32*). Variables on a path between two other variables are equivalent to mediating/moderating variables in statistics. For instance, the Bayesian network identified certain paths systematically leading to a higher probability of manuscript acceptance: While, as expected, the highest path coefficients for the prediction of an acceptance were the review score, a higher proportion of women as referees in interaction with a high proportion of women as authors also predicted whether a manuscript was accepted.

Table 2. Logistic mixed-effects models on the final editorial decision (accept) by field of research using the gender ratio as predictor. Mean estimate, 95% CI, and Bayes factor (β > 0) are reported for each variable.

Variable	Biomedicine and health science	Life science	Physical science	Social science
(Intercept)	−6.224	−4.698	−7.069	−5.124
	[−6.629, −5.827]	[−6.048, −3.366]	[−7.970, −6.174]	[−6.071, −4.200]
	1:20,000	1:20,000	1:20,000	1:20,000
Women proportion (authors)	0.129	0.050	0.205	−0.065
	[0.022, 0.235]	[−0.143, 0.244]	[0.115, 0.296]	[−0.291, 0.156]
	103:1	2:1	20,000:1	1:2
Women proportion (referees)	−0.154	−0.042	−0.041	−0.234
	[−0.240, −0.070]	[−0.206, 0.122]	[−0.119, 0.036]	[−0.448, −0.020]
	1:2,856	1:2	1:6	1:59
Review score	6.020	6.176	6.095	5.823
	[5.907, 6.134]	[5.936, 6.416]	[5.996, 6.194]	[5.470, 6.181]
	20,000:1	20,000:1	20,000:1	20,000:1
Agreement	1.214	0.667	0.708	0.202
	[1.086, 1.339]	[0.449, 0.879]	[0.613, 0.801]	[−0.122, 0.525]
	20,000:1	20,000:1	20,000:1	8:1
IF	−0.059	−0.140	0.058	−0.143
	[−0.112, −0.004]	[−0.215, −0.065]	[0.020, 0.095]	[−0.403, 0.114]
	1:57	1:20,000	832:1	1:6
Number of authors	0.002	−0.039	0.045	0.014
	[−0.006, 0.011]	[−0.053, −0.025]	[0.035, 0.054]	[−0.026, 0.055]
	2:1	1:20,000	20,000:1	3:1
Number of referees	−0.184	−0.160	−0.103	−0.300
	[−0.226, −0.142]	[−0.234, −0.0986]	[−0.133, −0.072]	[−0.420, −0.180]
	1:20,000	1:19,999	1:20,000	1:20,000
PR type: single-blind	0.532	0.117	1.185	1.091
	[0.97, 0.962]	[−1.228, 1.472]	[0.281, 2.110]	[−0.391, 2.592]
	105:1	1:1	162:1	14:1
Number of revision rounds	4.094	3.670	3.99	3.756
	[4.037, 4.152]	[3.578, 3.766]	[3.95, 4.04]	[3.624, 3.889]
	20,000:1	20,000:1	20,000:1	20,000:1
Sensitivity	0.93	0.93	0.93	0.92
Specificity	0.96	0.95	0.97	0.97

Tables S10 and S11 show further statistical tests on some interactions shown in the Bayesian network. We found that manuscripts written by women received better reviews when reviewed by other women in all scientific fields, although the effect was weak in case of manuscripts submitted to journals in life sciences. Manuscripts by women generally received worse reviews in social science journals using single-blind peer review (see table S10), but these journals are the minority in a field typically dominated by double-blind peer review. We also examined whether manuscripts written by women needed to be of higher quality to be published, by checking whether there was a negative interaction effect between authors' gender and the review score on the editorial decision. We found that such an interaction exists only in case of journals in biomedical and health sciences, while we found only weak effects in the case of journals in social sciences (see table S11).

Although we could not directly estimate the intrinsic quality of manuscripts (if this were possible even only in theory), we used the recommendations of referees as a control variable of the quality and used it to identify bias in the editorial decision. Our results indicated no statistical gender gap in acceptance rates. The Bayesian-learning model found that, after controlling for all other variables (including the recommendations), manuscripts by women were more likely to be accepted in journals of all disciplines except social sciences, where we did not find any significant gender difference. To quantify the effect of gender, we used the model to predict the final acceptance of all manuscripts in our dataset with the hypothetical scenario that all

Table 3. Logistic mixed-effects models on the final editorial decision (accept) by field of research using the first and last author's gender as predictors. Mean estimate, 95% CI, and Bayes factor ($\beta > 0$) are reported for each variable.

Variable	Biomedicine and health science	Life science	Physical science	Social science
(Intercept)	−6.116	−4.502	−7.020	−5.291
	[−6.530, −5.700]	[−5.844, −3.156]	[−7.960, −6.088]	[−6.282, −4.322]
	1:20,000	1:20,000	1:20,000	1:20,000
First author woman	0.001	−0.099	0.099	−0.065
	[−0.067, 0.069]	[−0.218, 0.022]	[0.035, 0.163]	[−0.259, 0.127]
	1:1	1:18	768:1	1:3
Last author woman	−0.056	−0.050	−0.034	0.039
	[−0.125, 0.014]	[−0.181, 0.081]	[−0.109, 0.024]	[−0.148, 0.223]
	1:16	1:3	1:8	2:1
Women proportion (referees)	−0.135	−0.063	−0.033	−0.190
	[−0.233, −0.037]	[−0.254, 0.130]	[−0.132, 0.066]	[−0.429, 0.044]
	1:302	1:3	1:3	1:16
Review score	6.017	6.246	6.056	5.785
	[5.889, 6.145]	[5.966, 6.532]	[5.928, 6.186]	[5.393, 6.181]
	20,000:1	20,000:1	20,000:1	20,000:1
Agreement	1.207	0.635	0.646	0.353
	[1.063, 1.353]	[0.387, 0.886]	[0.523, 0.769]	[−0.003, 0.710]
	20,000:1	20,000:1	20,000:1	38:1
IF	−0.059	−0.139	0.044	−0.173
	[−0.120, 0.002]	[−0.223, −0.053]	[−0.003, 0.091]	[−0.454, 0.113]
	1:33	1:1,817	28:1	1:8
Number of authors	0.005	−0.045	0.051	0.024
	[−0.005, 0.015]	[−0.061, −0.029]	[0.039, 0.063]	[−0.020, 0.068]
	6:1	1:20,000	20,000:1	6:1
Number of referees	−0.188	−0.199	−0.137	−0.286
	[−0.236, −0.141]	[−0.285, −0.114]	[−0.177, −0.098]	[−0.416, −0.155]
	1:20,000	1:20,000	1:20,000	1:20,000
PR type: single-blind	0.537	0.099	1.336	1.094
	[0.113, 0.974]	[−1.245, 1.435]	[0.405, 2.284]	[−0.406, 2.601]
	143:1	1:1	391:1	14:1
Number of revision rounds	4.100	3.707	4.018	3.834
	[4.036, 4.165]	[3.597, 3.819]	[3.961, 4.076]	[3.687, 3.988]
	20,000:1	20,000:1	20,000:1	20,000:1
Sensitivity	0.93	0.93	0.93	0.92
Specificity	0.97	0.96	0.97	0.96

authors were either men or women. In case of biomedical and health sciences journals, manuscripts written by women were predicted to be 5% more likely to be accepted than manuscripts written by men (women were predicted to be accepted in 45% of cases). While in the case of life and physical sciences journals, this probability decreased to 1.5% (for women, the prediction was 53% in both fields), in the case of social sciences journals, the probability was close to zero (with a predicted overall acceptance of 38% of manuscripts). This suggests that women are treated less favorably in the field of research where the ratio of women among authors is the highest (38% in social sciences versus 19% in physical sciences). Figure 3 shows the predicted editor decisions by authors' gender, controlling for different review scores. Last, the Bayesian-learning network further confirmed a systematic effect of gender on the match of authors and referees.

Given that peer review typically includes multiple rounds of revision, we also looked at the extent to which the length of the revision process could be influenced by the gender of authors and

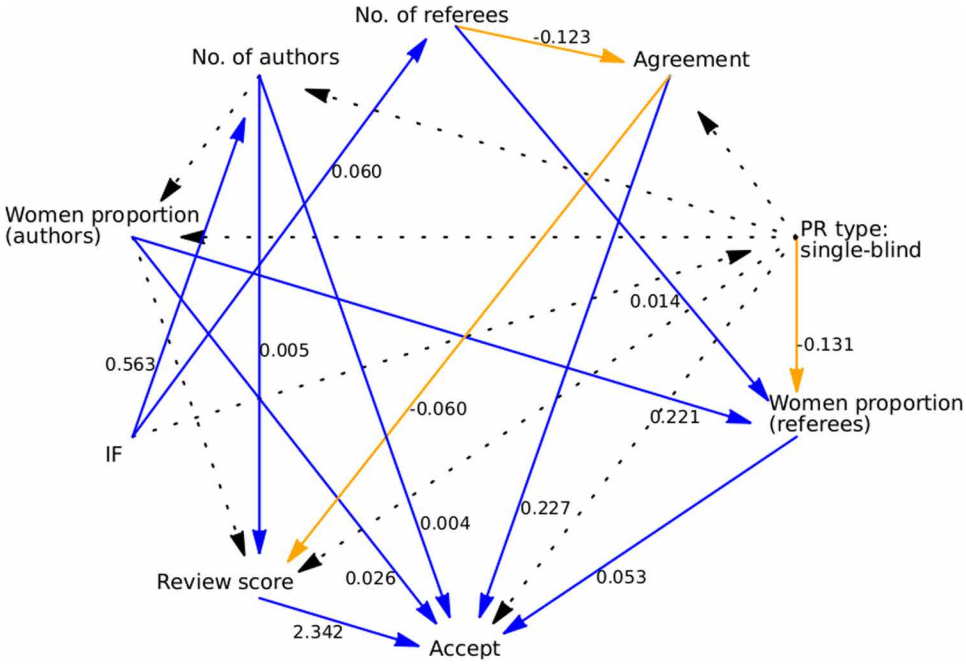

Fig. 2. Learned structure of the Bayesian network. For the sake of readability, we did not report the scientific field effect, which was linked to all nodes. Orange arrows indicate a negative relationship, and blue arrows indicate a positive relationship (dotted black, if the sign depends on the scientific field taken into consideration). Path coefficients are only shown for paths that were consistent across scientific fields. All path coefficients can be found in table S9.

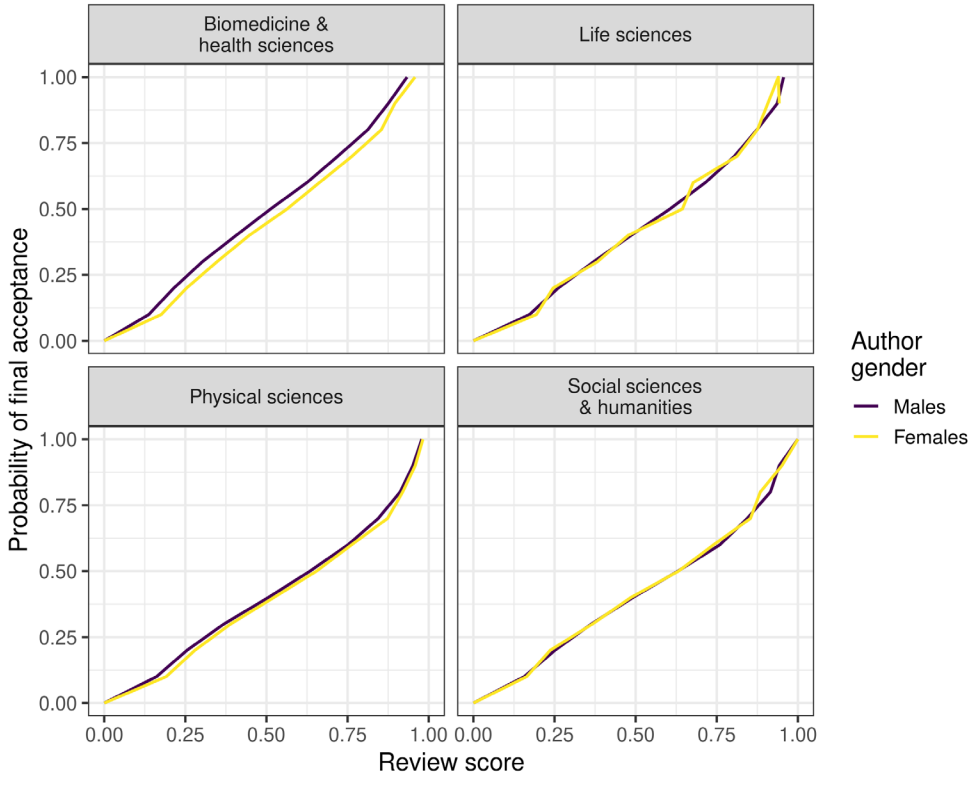

Fig. 3. Bayesian network predictions of the rejection probability by author gender, referee recommendation score panels, and field of research.

referees. Table 4 shows the estimates of a Poisson regression, which predicted the number of revision rounds that any manuscript eventually underwent before publication. We did not find any effect of gender on the number of required rounds of revision before publication. With the exception of journals in social sciences, the more women among the reviewer team, the higher the probability of more rounds of revisions before publication.

CONCLUSIONS

Although we could not perform a large-scale, multi-journal randomized experiment and worked only on existing journal data, our findings indicate that manuscripts submitted by women or coauthored by women are generally not penalized during the peer review process. We found that manuscripts by all women or cross-gender teams of authors had even a higher probability of success in many cases. This is especially so in journals in biomedicine, health, and physical sciences, thereby confirming previous research (*16*, *18*, *22*).

However, given that we did not have an objective or predefined estimation of the quality of manuscripts (if any) and could use only referee recommendations as an indication, this positive inclination by referees and editors could simply reflect some intrinsic characteristics of the manuscripts. Previous research suggests that women could be inclined to invest more in their manuscripts to prevent expected editorial bias (*10*, *33*), which could also explain why they submit fewer manuscripts (*18*, *23*, *24*, *28*). In this respect, the fact that manuscripts by cross-gender teams of authors received systematically more positive treatments in our sample could even reveal an exploitation opportunity by men, who benefit from collaborating with women colleagues.

Unfortunately, while the potential positive effect of higher inclusion of women in scientific networks has also been found in other studies (*10*, *34*), our dataset did not permit us to control any possible distortions in the potential pool of authors and referees available in each journal, age cohorts, or other (institutional/personal) status characteristics. Therefore, it is impossible to understand whether

Table 4. Poisson regression model predicting the number of rounds of reviews before manuscript's acceptance. Mean estimate, 95% CI, and Bayes factor ($\beta > 0$) are reported for each variable.

Variable	Biomedicine and health science	Life science	Physical science	Social science
(Intercept)	0.571	−1.171	−1.537	−1.427
	[0.488, 0.654]	[−1.444, −0.894]	[−1.821, −1.255]	[−1.710, −1.149]
	20,000:1	1:20,000	1:20,000	1:20,000
Women proportion (authors)	−0.002	−0.001	0.016	−0.006
	[−0.031, 0.027]	[−0.072, 0.071]	[−0.019, 0.052]	[−0.080, 0.069]
	1:1	1:1	4:1	1:1
Women proportion (referees)	0.037	0.083	0.049	0.007
	[0.013, 0.060]	[0.022, 0.143]	[0.019, 0.079]	[−0.069, 0.082]
	951:1	307:1	951:1	1:1
Review score	−0.389	1.712	1.812	2.251
	[−0.423, −0.355]	[1.642, 1.783]	[1.783, 1.842]	[2.153, 2.349]
	1:20,000	20,000:1	20,000:1	20,000:1
IF	−0.007	0.021	0.036	0.026
	[−0.021, 0.007]	[−0.002, 0.043]	[0.022, 0.049]	[−0.06, 0.111]
	1:6	28:1	20,000:1	3:1
Number of authors	0.002	0.010	0.014	−0.005
	[0, 0.005]	[0.004, 0.015]	[0.01, 0.017]	[−0.018, 0.008]
	43:1	6,666:1	20,000:1	1:3
Number of referees	0.053	0.065	0.106	0.089
	[0.042, 0.063]	[0.039, 0.091]	[0.095, 0.117]	[0.051, 0.127]
	20,000:1	20,000:1	20,000:1	20,000:1
Agreement	−0.022	0.031	0.056	0.033
	[−0.055, 0.012]	[−0.048, 0.111]	[0.019, 0.092]	[−0.075, 0.145]
	1:9	4:1	799:1	3:1
PR type: single-blind	−0.072	−0.051	0.095	−0.133
	[−0.155, 0.011]	[−0.303, 0.196]	[−0.192, 0.383]	[−0.554, 0.278]
	1:22	1:2	3:1	1:3

these potentially positive effects penalize older women and/or authors from less prestigious institutions (*14*). This also applies to the gender matching of authors and referees, which is in line with previous research (*13*). Rather than reflecting any editorial bias, this could simply reveal a gendered concentration of expertise in specific fields or a downstream effect of gendered patterns of citations (e.g., women/men authors citing in their manuscripts more references from women/men, who are possibly used by editors for referee selection).

It is worth noting that besides the lack of an objective measure of the quality of manuscripts, which is problematic and probably even impossible to establish consistently across fields, there are potentially important factors that are not included in our dataset. Some of them could be at least potentially minimized with extensive data search, such as the effect of authors' academic affiliation; others are impossible to capture, such as the role of authors' seniority and reputation, especially considering the scale and the cross-discipline nature of our dataset. For instance, it is extremely difficult to estimate the gender composition of various communities to calculate the potential pool of authors and referees in each journal, while we do not have robust proxies of authors' investment in manuscripts to estimate gender differences in submissions and volume of output (*23*).

In any case, our findings do not mean that peer review and journals are free from biases. For instance, the reputation of certain authors and the institutional prestige of their academic affiliation, not to mention authors' ethnicity or the type of research submitted, could influence the process, and these factors could also have gender implications (*30*, *35*). Here, data on the demographic composition of each disciplinary community and data on the invitation and acceptance to review at the journal level could help to complete our picture. On the other hand, these distortions could reflect built-in gendered norms and expectations, which could then persist and be reproduced either consciously or not, even when their expected "true" effects have disappeared (*33*). Considering the persistent and usually non-acknowledged obstacles that women still face in hypercompetitive academia (*36*), these expectations would be consistent even if the editorial processes of a set of journals were not objectively biased against women (*24*).

Our findings suggest that promoting more gender diversity in editorial teams and pools of referees could help scholarly journals to inform potential authors and referees about their attention to these factors and to stimulate the inclusion and participation of women (*24*, *37*, *38*). While diversity is beneficial for science and innovation per se (*37*), in this case, it would also be a signal that could contribute to reshaping the social construction of gender categories in academia and help scholarly journals to increase submission rates by women. Unfortunately, our research could not examine these complex expectations and norms characterizing academic life across all its spectrum, including academic choices of priorities and specialties (*5*, *39*), and educational stereotypes (*40*).

As previously stated, our aim was to concentrate on peer review, which is an important process determining the quantity and prestige of scholars' publication, while contributing to shaping their reputation in the community. However, studies capable of combining academic standards of promotion and the effect of author prestige and institutional affiliation on editorial process in scholarly journals are required to examine the complex nexus of gender discrimination (and even other sources of bias) in academia (*33*), including reconstructing the gender gap–gender bias link in a comprehensive manner. However, this raises the problem of data availability (*26*). While data sharing on editorial processes of journals should be encouraged more systematically on a large scale with collaboration between publishers and independent research groups (*27*, *41*, *42*), examining structural mechanisms that determine academic opportunities requires data integration from various sources (i.e., funding agencies, academic institutions, and scholarly citation databases). Only collaborative efforts on data sharing by various stakeholders will help us to grasp all the pieces of this gender puzzle.

MATERIALS AND METHODS

Data overview

Our dataset included internal data for 157 scholarly journals between 2010 and 2016, of which 61 were in biomedicine and health, 50 in physical sciences (including engineering and computer science), 24 in life sciences, and 22 in social sciences and humanities. Details on journal selection and the protocol for data sharing are provided in the Supplementary Materials. Data consisted of all actions or events performed by one of the journal editors, such as inviting referees, receiving reviews, or deciding about manuscripts. They included 753,909 submitted manuscripts, of which 389,431 (51.7%) were sent out to referees.

To ensure better comparability of peer review and editorial standards, in our analyses, we only considered journals included in the Journal Citation Report based on the Web of Science (WoS) and with an impact factor (98% of our observations, see fig. S1).The resulting dataset included 145 journals and 348,223 submissions. Because of a few missing observations in the data, the actual numbers of complete observations used in the analysis were 348,118 (Table 1). These included a total of 1,689,944 authors and 745,693 referees, with an average of 2.1 completed reviews per manuscript.

The dataset includes the following variables: Manuscript ID, unique manuscript identifier; SubmissionDate, initial submission date; JournalID, unique journal identifier; ScientificArea, journal's field of research (scientific area); PRType, peer review type; IFRounded, journal's impact factor rounded to integer (this was to ensure journal's anonymity); nAuthors, number of authors; NumRounds, number of review rounds; Agreement, referee agreement score; nRev, number of referees; RevScore, review score; AutRatFem, ratio of women authors; RevRatFem, ratio of women referees; FirstAuthorGender, gender of the first author; LastAuthorGender, gender of the last author; FinalDecision, final editorial decision.

The number of manuscripts reviewed by these journals was approximately constant over time, with about 50,000 editorial decisions per year, and a majority of records from physics and biomedicine and health journals (see fig. S2). Given that we aimed to focus on the peer review process, we considered each submitted manuscript as our unit of analysis. Statistics showed that the proportion of accepted papers varied across scientific fields, from 51.9% in life sciences to 37.7% in social sciences (see fig. S3).

Referee recommendations were combined so that a review and an agreement score were calculated for each manuscript (*29*). While in (*29*) the former was bounded in the [0,1] interval, we multiplied these with 100 to make estimates in the table more informative. The review score was calculated independently of the number of referees, with higher values reflecting more positive referee recommendations. Following (*29*), the agreement score was calculated in the same interval, with higher values meaning a stronger agreement between referee recommendations (*29*).

More specifically, to calculate review scores, we first recoded each referee recommendation (which sometimes appeared as nonstandard expressions in our database) in a standard ordinal scale: reject, major revisions, minor revisions, accept. We then derived the set of all possible unique combinations of recommendations for each manuscript (from now on, the "potential recommendation set"). Using this set, we counted the number of combinations that were clearly less favorable (#worse) or more favorable (#better) than that actually received by the manuscript (e.g.,{accept, accept} was clearly better than {reject, reject}). Last, we calculated the score of each manuscript as follows

$$\text{reviewScore} = \frac{\#\text{worse}}{\#\text{better} + \#\text{worse}} \quad (1)$$

Note that while (*29*) calculated a disagreement score, here, we assumed an agreement score for each manuscript, i.e., one minus the number of referee recommendations that should be changed to reach a perfect agreement between referees divided by the number of referees assigned to the manuscript. This permitted full comparability between manuscripts receiving a different number of reviews.

Statistical analysis

We estimated our mixed-effects models using the R 3.6.1 platform (*43*). Our plots were generated using the ggplot2 package on the same platform. In all linear and logistic mixed-effect models, we included random effects for journals. We tested all model specifications including nested random effects for journals by considering the potential distortions due to sampling by publishers and found no effect on results. To comply with the data sharing protocol, we did not report details here to avoid journal identification. Mixed-effects models were estimated using the brms package (*44*) and are the outcome of four independent chains, each including 10,000 iterations (5000 burn-in + 5000 sampling). To ensure that the estimates are reliable, we checked that all scale reduction factors ($\hat{R}$) (*45*) were below 1.01. In each table, we reported the coefficients' mean estimates, 95% credible intervals (CIs), and the Bayes factor corresponding to the hypothesis $\beta > 0$. The interpretation of Bayes factors was done following the recommendations in (*46*). To compute the proportion of variance explained by the models (pseudo-R^2), we used the approach proposed in (*47*). All models used flat priors with a zero mean for all model parameters.

Bayesian network

Our analysis followed a previous study on network effects on peer review in four journals (*29*). Building a Bayesian network was pivotal in modeling complex interactions between variables and potential indirect paths of bias (*31*). We selected this method over alternative machine learning techniques (e.g., neural networks) as it allowed us to generate a directed acyclic graph that was more appropriate to examine the structure of relations characterizing the editorial process. Furthermore, this graph permitted us to calculate the probability of an event (e.g., a rejection) depending on the value of other variables of interest (e.g., all authors being men).

The Bayesian network was estimated using the bnlearn package. We first trained the network on a random sample of 80% of all available manuscripts, while the other 20% were used as independent test data for model validation. Note that all nodes corresponded to the variables used in the statistical models presented in the main text. The structure of the Bayesian network and the direction of influence were learned through various constraint- and score-based structure learning algorithms. All algorithms resulted in structurally similar graphs, which were then aggregated in one network by including all links learned by at least 70% of structure learning algorithms. Figure 2 shows the resulting network. Note that we only imposed restrictions on the structure learning algorithms such that links pointing from the review score and the editorial decision to any of the other nodes were not allowed, as were any links that were chronologically impossible.

It is worth noting here that our data were imbalanced in respect of certain variables considered in the Bayesian network. This is the case of the lower amounts of women among submission authors and the overrepresentation of manuscripts from physical sciences. On the one hand, this, in principle, implies that the learned structure of the network cannot be fully generalized to all manuscripts. However, all model diagnostics showed that these imbalances did not affect our results (see table S6). Therefore, we decided not to rebalance data manually, which would have been difficult given the amount of variables characterizing our dataset and, in any case, would have led to loss of information.

Gender guessing

The method used for gender guessing was inspired by previous research (*1*, *13*, *48*) and prioritized accuracy above other considerations (*49*). We followed a standard disambiguation algorithm recently validated on a dataset of scientist names extracted from the WoS database and tested with the same time window used in our study (*50*).

Gender was estimated for each individual record following a multistage gender inference procedure consisting of three steps, in order of priority. First, we performed preliminary gender guessing using, when available, gender salutation (i.e., Mr., Mrs., Ms., etc.). Second, we queried the Python package gender-guesser about the extracted first names and country of origin, if any, to corroborate our procedure. To maximize accuracy, we did not follow gender-guesser for names classified as mostly_man, mostly_woman, andy (androgynous), or unknown (name not found). Previous research shows that gender-guesser achieves the lowest misclassification rate and minimizes bias (*50*). We then queried the best performer gender inference service, Gender API (https://gender-api.com/), and used the returned gender whenever we found a minimum of 62 samples with at least 57% accuracy. These confidence parameters for Gender API permitted us to comply with the optimal values ensuring that the rate of misclassified names did not exceed 5% [see Benchmark 2 in (*50*)].

As a result, we were able to estimate the gender of 82% of referees and 77% of authors (table S7). The remaining scientists were assigned an unknown gender, a proportion that is in line with up-to-date nonclassification rates for names of scientists found in literature (*50*). This method is robust because it implies that a human coder would hardly be able to identify these uncertain gender cases, thereby potentially introducing further bias, if involved.

Our three-step gender guessing procedure was mostly based on gender-guesser (table S8), which is currently the best tool to assign names by origin. We estimated gender of 57% of authors and 63% of referees from this library, which also showed a fraction of misclassification under 5% [see table 6 in (*50*)]. Note that the validation performed by (*50*) limited misclassification to 1.5% for European names, 3.6% for African names, and 6.4% for Asian names [see table 5 in (*50*)]. We followed Gender API to assign the gender to 13% of

referees and 16% of authors. The percentage of misclassification of this gender service was 2.1% for European names, 4.7% for African names, and 11.2% for Asian names [see table S5 in (*50*)]. Last, salutation was used to guess gender of 4% authors and 6% referees.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/2/eabd0299/DC1

REFERENCES AND NOTES

1. V. Larivière, C. Ni, Y. Gingras, B. Cronin, C. R. Sugimoto, Bibliometrics: Global gender disparities in science. *Nature* **504**, 211–213 (2013).
2. J. R. Cole, H. Zuckerman, The productivity puzzle: Persistence and change in patterns of publication of men and women scientists. *Adv. Motiv. Achiev.* **2**, 217–258 (1984).
3. M. L. Dion, J. L. Sumner, S. McLaughlin Mitchell, Gendered citation patterns across political science and social science methodology fields. *Polit. Anal.* **26**, 312–327 (2018).
4. R. Jagsi, E. A. Guancial, C. C. Worobey, L. E. Henault, Y. Chang, R. Starr, N. J. Tarbell, E. M. Hylek, The "gender gap" in authorship of academic medical literature—A 35-year perspective. *NEJM* **355**, 281–287 (2006).
5. J. D. West, J. Jacquet, M. M. King, S. J. Correll, C. T. Bergstrom, The role of gender in scholarly authorship. *PLOS ONE* **8**, e66212 (2013).
6. P. van den Besselaar, U. Sandström, Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PLOS ONE* **12**, e0183301 (2017).
7. P. van Arensbergen, I. van der Weijden, P. van den Besselaar, Gender differences in scientific productivity: A persisting phenomenon? *Scientometrics* **93**, 857–868 (2012).
8. E. Z. Cameron, A. M. White, M. E. Gray, Solving the productivity and impact puzzle: Do men outperform women, or are metrics biased? *Bioscience* **66**, 245–252 (2016).
9. J. Huang, A. J. Gates, R. Sinatra, A.-L. Barabási, Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 4609–4616 (2020).
10. D. L. Teele, K. Thelen, Gender in the journals: Publication patterns in political science. *PS Polit. Sci. Polit.* **50**, 433–447 (2017).
11. A. Akbaritabar, F. Squazzoni, Gender patterns of publication in top sociological journals. *Sci. Technol. Hum. Values* , (2020).
12. K. Weisshaar, Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Soc. Forces* **96**, 529–560 (2017).
13. M. Helmer, M. Schottdorf, A. Neef, D. Battaglia, Research: Gender bias in scholarly peer review. *eLife* **6**, e21718 (2017).
14. J. Lerback, B. Hanson, Journals invite too few women to referee. *Nature* **541**, 455–457 (2017).
15. C. J. Lee, C. R. Sugimoto, G. Zhang, B. Cronin, Bias in peer review. *J. Am. Soc. Inf. Sci. Technol.* **64**, 2–17 (2013).
16. R. M. Borsuk, L. W. Aarssen, A. E. Budden, J. Koricheva, R. Leimu, T. Tregenza, C. J. Lortie, To name or not to name: The effect of changing author gender on peer review. *Bioscience* **59**, 985–989 (2009).
17. G. Østby, H. Strand, R. Nordås, N. P. Gleditsch, Gender gap or gender bias in peace research? publication patterns and citation rates for journal of peace research, 1983–2008. *Int. Stud. Perspect.* **14**, 493–506 (2013).
18. C. L. Tudor, D. J. Yashar, Gender and the editorial process: *World Politics*, 2007–2017. *PS Polit. Sci. Polit.* , 1–11 (2018).
19. E. Grossman, A gender bias in the European Journal of Political Research? *Eur. Polit. Sci.* **19**, 416–427 (2020).
20. D. Card, S. D. Vigna, P. Funk, N. Iriberri, Are referees and editors in economics gender neutral? *Q. J. Econ.* **135**, 269–327 (2020).
21. C. W. Fox, C. E. T. Paine, Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecol. Evol.* **9**, 3599–3619 (2019).
22. C. W. Fox, C. S. Burns, J. A. Meyer, Editor and reviewer gender influence the peer review process but not peer review outcomes at an ecology journal. *Funct. Ecol.* **30**, 140–153 (2016).
23. E. Hengel, Publishing while female, in *Women in Economics*, S. Lundberg, Ed. (CEPR Press, 2020), pp. 80–90.
24. N. E. Brown, Y. Horiuchi, M. Htun, D. Samuels, Gender gaps in perceptions of political science journals. *PS Polit. Sci. Polit.* **53**, 114–121 (2020).
25. H. A. Edwards, S. Julia, H. L. Dugdale, Gender differences in authorships are not associated with publication bias in an evolutionary journal. *PLOS ONE* **14**, e0217251 (2019).
26. F. Squazzoni, P. Ahrweiler, T. Barros, F. Bianchi, A. Birukou, H. J. J. Blom, G. Bravo, S. Cowley, V. Dignum, P. Dondio, F. Grimaldo, L. Haire, J. Hoyt, P. Hurst, R. Lammey, C. M. Callum, A. Marušić, B. Mehmani, H. Murray, D. Nicholas, G. Pedrazzi, I. Puebla, P. Rodgers, T. Ross-Hellauer, M. Seeber, K. Shankar, J. Van Rossum, M. Willis, Unlock ways to share data on peer review. *Nature* **578**, 512–514 (2020).
27. F. Squazzoni, F. Grimaldo, A. Marusic, Publishing: Journals could share peer-review data. *Nature* **546**, 352 (2017).
28. T. König, G. Ropers, Gender and editorial outcomes at the *American Political Science Review*. *PS Polit. Sci. Polit.* **51**, 849–853 (2018).
29. G. Bravo, M. Farjam, F. Grimaldo Morenoo, A. Birukou, F. Squazzoni, Hidden connections: Network effects on editorial decisions in four computer science journals. *J. Informet.* **12**, 101–112 (2018).
30. A. Tomkins, M. Zhang, W. D. Heavlin, Reviewer bias in single- versus double-blind peer review. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12708–12713 (2017).
31. N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers. *Mach. Learn.* **29**, 131–163 (1997).
32. K. B. Korb, A. E. Nicholson, The causal interpretation of Bayesian networks, in *Innovations in Bayesian Networks* (Springer, 2008), pp. 83–116.
33. S.-J. Leslie, A. Cimpian, M. Meyer, E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines. *Science* **347**, 262–265 (2015).
34. L. G. Campbell, S. Mehtani, M. E. Dozier, J. Rinehart, Gender-heterogeneous working groups produce higher quality science. *PLOS ONE* **8**, e79147 (2013).
35. L. Holman, D. Stuart-Fox, C. E. Hauser, The gender gap in science: How long until women are equally represented? *PLOS Biol.* **16**, e2004956 (2018).
36. S. Lundberg, J. Stearns, Women in economics: Stalled progress. *J. Econ. Perspect.* **33**, 3–22 (2019).
37. M. W. Nielsen, C. W. Bloch, L. Schiebinger, Making gender diversity work for scientific discovery and innovation. *Nat. Hum. Behavior* **2**, 726–734 (2018).
38. M. R. Berenbaum, Speaking of gender bias. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8086–8088 (2019).
39. E. Cech, B. Rubineau, S. Silbey, C. Seron, Professional role confidence and gendered persistence in engineering. *Am. Sociol. Rev.* **76**, 641–666 (2011).
40. S. J. Ceci, W. M. Williams, Understanding current causes of women's underrepresentation in science. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3157–3162 (2011).
41. G. Pinholster, Journals and funders confront implicit bias in peer review. *Science* **352**, 1067–1068 (2016).
42. G. Bravo, F. Grimaldo, E. López-Iñesta, B. Mehmani, F. Squazzoni, The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nat. Commun.* **10**, 322 (2019).
43. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2018).
44. P.-C. Bürkner, brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80**, 1 (2017).
45. A. Gelman, D. B. Rubin, Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
46. R. E. Kass, A. E. Raftery, Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
47. A. Gelman, B. Goodrich, J. Gabry, A. Vehtari, R-squared for Bayesian regression models. *Am. Stat.* **73**, 307–309 (2019).
48. F. Karimi, C. Wagner, F. Lemmerich, M. Jadidi, M. Strohmaier, Inferring gender from names on the web: A comparative evaluation of gender detection methods, in *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion (International World Wide Web Conferences Steering Committee, 2016), pp. 53–54.
49. J. Kim, J. Diesner, Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *J. Assoc. Inf. Sci. Technol.* **67**, 1446–1461 (2016).
50. L. Santamaría, H. Mihaljević, Comparison and benchmark of name-to-gender inference services. *PeerJ Comp. Sci.* **4**, e156 (2018).

Acknowledgments: We would like to thank the IT office staff of all partners for support on initial data extraction. The analysis was carried out exploiting the Linnaeus University Centre for Data Intensive Sciences and Applications high-performance computing facility. Access to data was possible thanks to the "PEERE Protocol for Data Sharing," co-signed by all involved partners on 1 March 2017. **Funding:** This work was supported by the TD1306 COST Action "New Frontiers of Peer Review." This work was also partially supported by the Spanish Ministry of Science, Innovation and Universities (MCIU), the Spanish State Research Agency (AEI), and the European Regional Development Fund (ERDF) under project RTI2018-095820-B-I00. A preliminary version of the manuscript received confidential comments by J. Marsh and A. Marengoni. **Author contributions:** F.S. designed the study and wrote and revised the manuscript. F.G. coordinated data collection and wrote and revised the manuscript. P.D. collected and prepared data. G.B. designed and performed the analysis and wrote and revised the manuscript. M.F. designed the analysis and wrote and revised the manuscript. A.M. contributed to the study design and wrote and revised the manuscript. M.W. and B.M. contributed to the study design, provided data, and revised the manuscript. A.B. contributed to the study design and revised the manuscript. **Competing interests:** B.M. declares a competing interest, being currently employed as Reviewer Experience Lead at Elsevier. A.B.

declares a competing interest, being currently employed as Executive Editor in the Computer Science team at Springer Nature. M.W. declares a competing interest, being currently employed as Researcher Advocate, Content Peer Review at John Wiley and Sons and having stock options in John Wiley & Sons, his employer. Neither of them had access to the database, elaborated any version of the dataset, or were involved in data analysis. The authors declare no other competing interests. **Data and materials availability:** Our dataset is made available at https://dataverse.harvard.edu/privateurl.xhtml?token=7b70ab08-b062-4589-b024-4584a130ed06 with all records required to rerun our analysis.

Submitted 27 May 2020
Accepted 13 November 2020
Published 6 January 2021
10.1126/sciadv.abd0299

Citation: F. Squazzoni, G. Bravo, M. Farjam, A. Marusic, B. Mehmani, M. Willis, A. Birukou, P. Dondio, F. Grimaldo, Peer review and gender bias: A study on 145 scholarly journals. *Sci. Adv.* **7**, eabd0299 (2021).