Gender Homophily: In-Group Citation Preferences and the Gender Disadvantage

Sifan ZHOU*
Xiamen University
Xiamen, China
zhousifan@gmail.com
* corresponding author

Sen CHAI
ESSEC Business School
Cergy-Pontoise, France
chai@essec.edu

Richard FREEMAN
Harvard University & NBER
Cambridge, MA
freeman@nber.org

Abstract
Forward citations are widely used to measure the scientific merit of articles and in turn the research productivity of scientists which have consequential economic implications to scientists' careers such as in hiring, promotion and retention. They are also an important measure of the subsequent usage and diffusion of the research produced. Using an extensive sample of articles from PubMeD, we find a significant gender gap in articles' forward citations, such that articles written by women receive fewer citations than those by men. By examining citations patterns received from men versus women separately, we find that forward citations exhibit gender homophily in that women receive significantly more citations from women while men receive significantly more citations from men. Delving deeper, we find that gender homophily in forward citations stem from the size and composition of networks. Given that the majority of life scientists are men and men publish more than women, gender homophily leaves women at a disadvantage.

Keywords
Gender Gap, Gender Homophily, Diffusion of Knowledge, Forward Citations, Collaborative Network

Gender Homophily: In-Group Citation Preferences and the Gender Disadvantage

Despite improved equity between men and women in education attainment (*1*, *2*), the gender gap is still pervasive in research productivity (*3–5*) – measured in academia by some function of publication and forward citation counts. The literature has however mainly focused on the gender gap in publication counts (*6–9*). Across various fields and time periods women publish less than men (*4*) – in fact, depending on the studied sample, the disadvantage ranges from 11% to 57% as a share of men's publications (*6*) (see **Fig 1A**). And since this gender gap in publishing cannot be fully explained by various factors (*6*, *10–12*) such as field, career length, academic rank, teaching burden, and research funding, it is usually referred to as a "productivity puzzle" (*7*).

Fewer studies investigate gender differences in forward citations as a measure of the scientific impact and subsequent usage and diffusion of the research produced (*13*, *14*). Most existing studies either focus on a specific narrow field or geography (*15–20*) or document the existence of a gender gap without digging into the mechanisms (*4*, *5*). Dworkin et al. (19) is a notable exception finding that men tend to cite men more. However, since it focuses solely on the (backward) citation decision made by citers it is hard to disentangle whether citations made to men versus women were based on quality differences in the articles or due to gender preference. Instead, we look at forward citations where we can compare similar papers by the two genders based on observables and explore how and whether they are treated differently. We, therefore, examine the gender of both the person making a citation and the one receiving citations. From the perspective of the person making citations, we examine the tendency of individuals to obtain citations from the same gender, and the extent to which such a tendency is associated with homophily of research collaboration networks. From the perspective of the person receiving citations, we examine how gender affects the way the individual's work is subsequently cited and used.

For our analysis, we combined two datasets: PubMed and Microsoft Academic Graph (MAG) (*21*), as bridged by MAG's existing matching. This allows us to take advantage of MAG's disambiguation of authors and forward citations, and PubMed's bibliometric data. We restrict our focal sample to US-based journal articles published in English between 2002 and 2017[1] without missing data and obtain a final sample of 2,432,806 publications (supplemental materials 1). Since 95% of journal articles in our sample have more than one author, we classify the gender of multi-authored articles as women-led[2] following authorship norms in the life sciences using either last author – usually the principal investigator, lab head or advisor – or first author – usually the main researcher or advisee. For each article, we use three-year forward citation windows[3]. Consistent with prior findings on gender, our data shows that women-led articles receive fewer forward citations than men-led articles (see **Figs 1B** and **1C** and supplemental materials 2). The absolute difference in forward citations by gender ranges from 1.06 to 1.92 fewer cites for women-led articles (see **Table S1**), which accounts for over 10% of the average forward citations per article irrespective of gender. This gap will also accumulate and magnify as an article ages and likely have significant impact on future career performance and survival.

To explain this gender gap in forward citations we turn to the sources of these citations and the gender of the person making the citation. We first account for observable factors identified in prior research that influence forward citations and exclude self-citations as prior works and our data show that men have a greater propensity to self-cite than women (*22–24*) (supplemental materials 3). There are additional factors that may account for this difference: article quality and a gender preference in citations, due possibly to network effects and possibly to some underlying gender bias. If article quality is the only determinant of citations, then the article's quality should be recognized analogously by potential citers irrespective of the author's gender. This however is not what we observe after excluding self-citations from the overall citations and dividing the remaining citations into those from male citers and female

[1] Our data from 2018-2020 are used to compute three-year forward citations for the 2017 focal articles.

[2] We also classify the gender of an article using the share of women authors, instead of women-led first or last author, and find robust results as shown in supplemental materials 6.

[3] We repeat all our analyses using five-year and ten-year citation windows and find robust results throughout.

citers. Instead, we find that men-led articles are consistently more likely to receive citations from men (blue circle line) than from women (red circle line); while women-led articles are more likely to receive citations from women (red square line) than from men (blue square line) as shown in **Figs 2A** and **2B**[4]. This implies that forward citations exhibit gender homophily. OLS regressions controlling for observables and excluding self-citations in **Figs 2C** reinforce these findings. The figure shows that women-led articles receive significantly more citations from other women compared to men-led articles, while women-led articles receive significantly less citations from other men compared to men-led articles.

Some of the difference may be due to women and men being in different fields or differences in the special topics women and men researchers choose within field. However, **Figs 2D** and **2E** show that gender homophily still holds when articles are separated into 288 fields. Indeed, in most fields women-led articles still receive a higher percentage of citations from women than men-led articles even in those where the share of women-led articles is low (bottom left corner). MAG also conducted topic modelling, and articles in our sample were tagged to 59,411 distinct scientific concepts (*25*). Our analysis indicates that articles written by each gender are still favored by citers of the same gender after controlling for concept fixed effect (see supplemental materials 4, **Tables S13** and **S14**). Given the estimated preferences for the same gender are similar, if women and men were equally represented among authors, the preferences would benefit neither in terms of numbers of citations received for papers. But women make up only 30% of authors in the life sciences (*6*), so they are at a disadvantage in garnering forward citations even with comparable gender preference.

What might explain gender homophily in citations? For one, men and women may maintain different professional networks such as who they collaborate with, who they talk to, who they sit with in conferences, etc. These networks will influence how knowledge diffuses and even research direction. We therefore use the size and composition of their recent (using three-year prior publications[5]) first-degree collaborators as proxy of their professional network (supplemental materials 5). We observe in **Fig 3A** that the collaborative networks of articles written by men are larger than those by women, and their gender composition also exhibit homophily consistent with prior findings (*26*, *27*), where both genders have a higher share of same-gender collaborators than predicted by randomness. The larger collaborative networks of men and the large male share of those collaborators should reinforce the observed gender homophily in citations, which is what we find in **Fig 3B**. Specifically, as the share of female collaborators increases citations from women-led articles also increase significantly, while citations from men-led articles decrease significantly.

To what extend could the gender homophily observed in collaborator networks and forward citations be viewed as gender bias, in that researchers have a preference to collaborate with and cite the same gender and against the opposite gender. If this were a general phenomenon, we would expect to find it outside the authors' immediate collaboration network where potential citers would infer the gender of an author through the author's first name and base their citation decision of an article on that inference. We do not observe such a bias in the data by building a subsample of authors whose articles sometimes report author first names and sometimes do not, but whose gender we have imputed from our disambiguated author database (supplemental materials 5). As shown in **Fig 3C**, we find after controlling for observables and individual fixed effects that citations from male citers to female articles not reporting first names are not significantly different from zero, and at the same time citations from female citers to female articles are also not significantly different from zero. This entails that for the same author, reporting first name or not (i.e. whether the gender could be guessed by strangers or not) does not change the degree of bias this author is subject to from both male and female citers. Thus, the source of the

---

[4]Note that an average article receives ~3 citations from male citers, but only ~1 citation from female citers. This is because there are fewer female citers than male citers.

[5]Here we use focal articles from 2005-2017 such that data from 2002-2004 are used to build the collaborator network for focal articles in 2005.

gender gap does not appear to be from outside authors' immediate collaboration networks but from within.

Taken together, our findings demonstrate that forward citations are influenced by gender homophily in forward citations that stem from network effects. Given that the majority of life scientists are men and men publish more than women, gender homophily leaves women at a disadvantage. Our analysis also has implications for how far information flows might be from ideal. Indeed, advancements in the scientific endeavor are enabled by "climbing on the shoulders of giants" (*28*), but our findings suggest that it is done so on the shoulders of more *similar* giants. This may lead to a certain degree of entrenchment and lack of diversity in how knowledge diffuses and subsequently recombined (*29*, *30*). This observation of homophily stemming from network effects can also be potentially generalized to researcher communities with dissimilar network sizes, such as big vs. small country, highly ranked research institutions vs. lower-ranked ones, etc.

Figure 1A. Share of women among life scientists, including biological, agricultural, environmental, and health sciences. (Details on data sources are found in supplemental material 1)

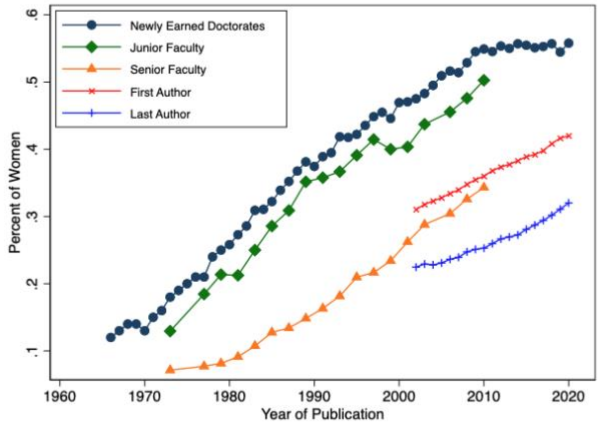


Figure 1B. Forward citations received, focal article gender is classified by *last* author

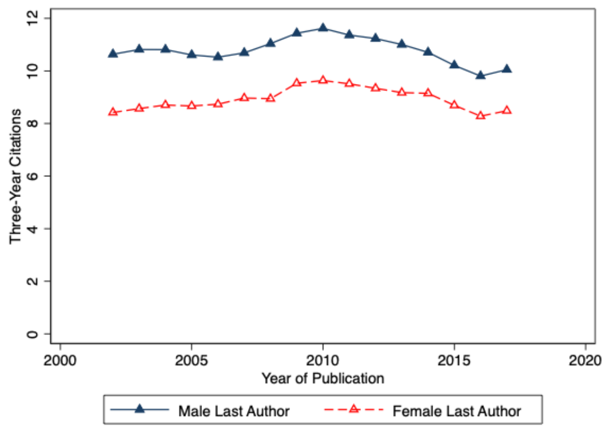


Figure 1C. Forward citations received, focal article gender is classified by *first* author

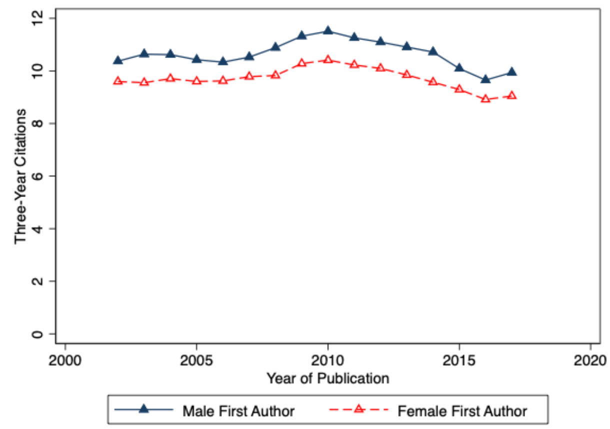


Figure 2A. Gender homophily in citations, forward citations made by *female* and *male* citers to *female* and *male* focal articles, focal article and citer gender are classified by *last* author

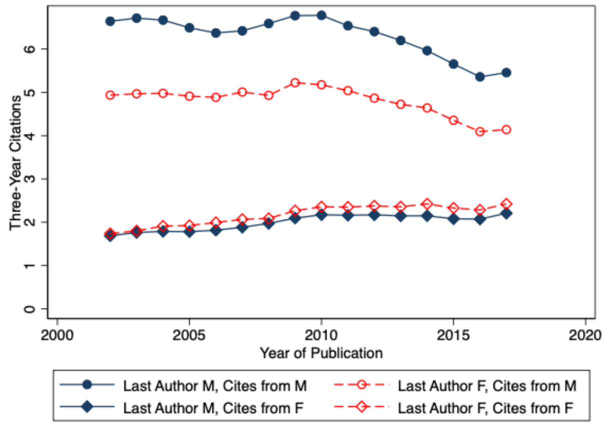


Figure 2B. Gender homophily in citations, forward citations made by *female* and *male* citers to *female* and *male* focal articles, focal article and citer gender are classified by *last* author

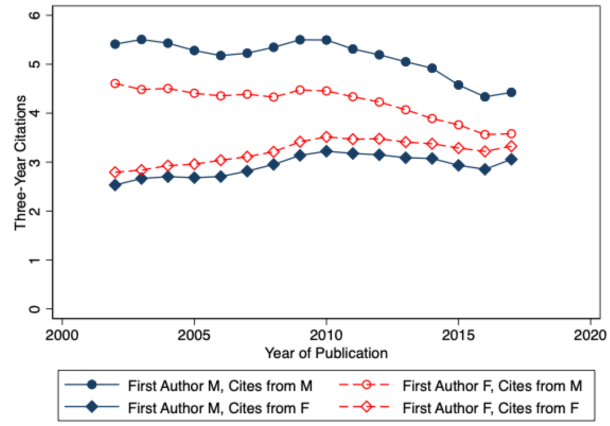


Figure 2C. Coefficients for citations received from *male* and *female* citers regressed on *female* focal articles, focal article and citer gender are classified by *last* and *first* author

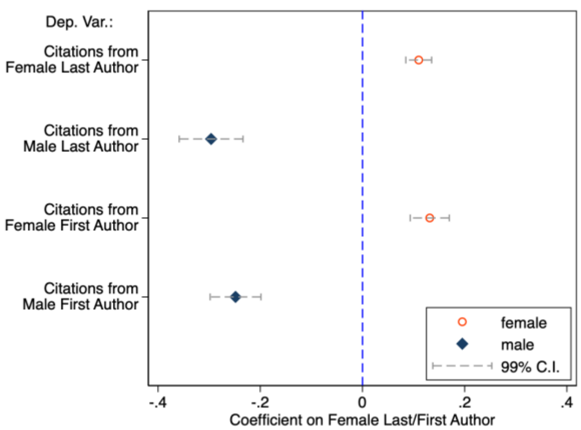


Figure 2D. Percent of forward citations made by *female* citers to *female* and *male* focal articles by field, focal article and citer gender are classified by *last* author (in fields with > 500 articles)

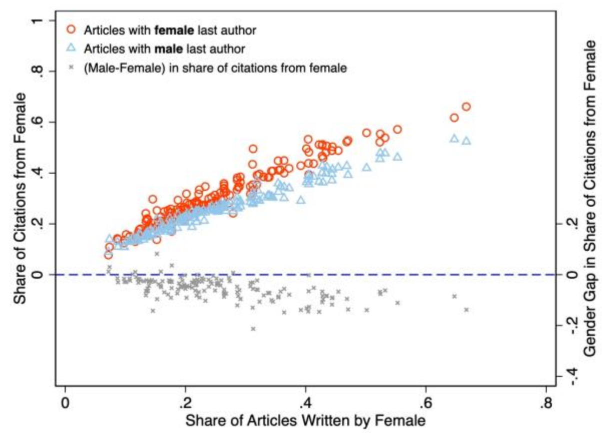


Figure 2E. Percent of forward citations made by *female* citers to *female* and *male* focal articles by field, focal article and citer gender are classified by *first* author (in fields with > 500 articles)

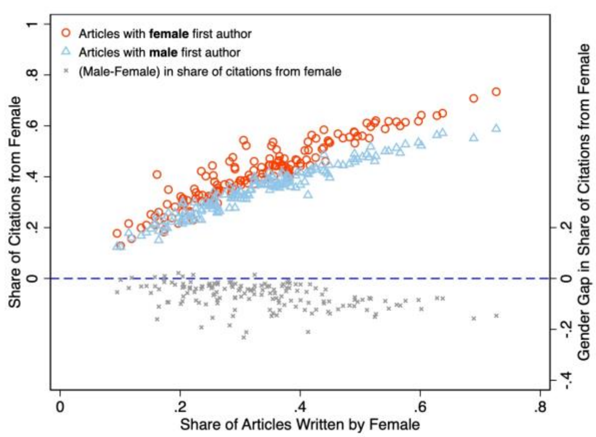


Figure 3A. Degree-one collaborator network, focal article gender is classified by *last* and *first* author

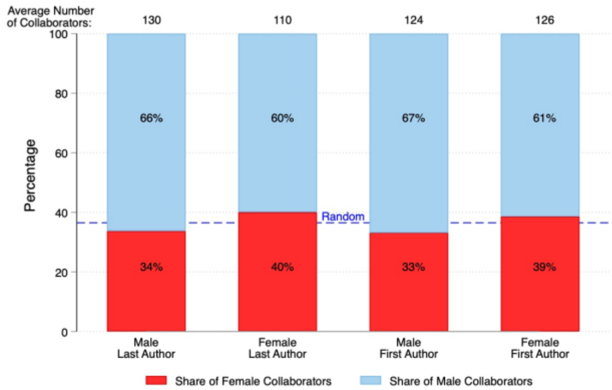


Figure 3B. Coefficients for citations received from *female* and *male* citers regressed on the female share of degree-one collaborators, focal article and citer gender are classified by *last* and *first* author

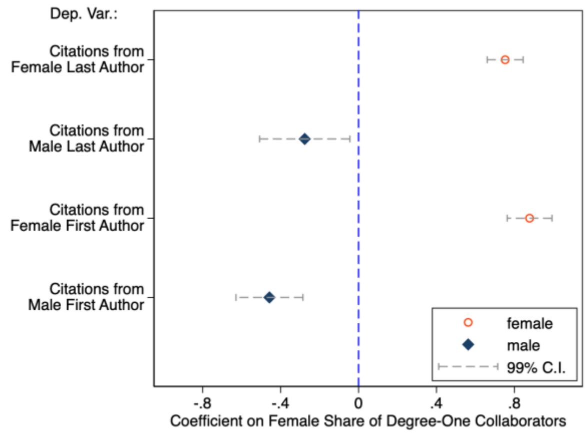
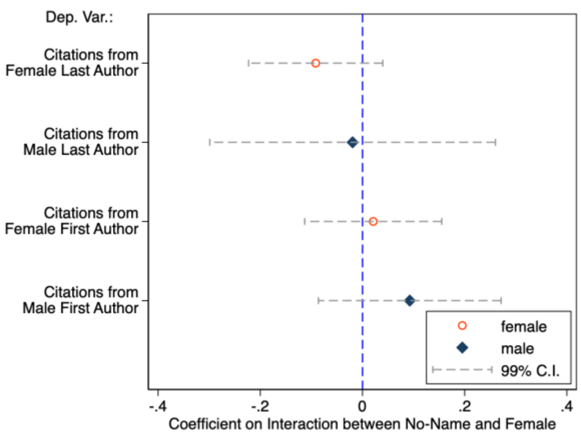


Figure 3C. Coefficients for citations received from *male* and *female* citers each regressed on the interaction between an article not displaying first names and written by *female* author, focal article and citer gender are classified by *last* and *first* author



References

1. OECD, *Education at a Glance 2012* (2012; https://www.oecd-ilibrary.org/content/publication/eag_highlights-2012-en).

2. D. Autor, M. Wasserman, Wayward sons: The emerging gender gap in labor markets and education. *Third Way Report*. **20013** (2013).

3. S. J. Ceci, W. M. Williams, Understanding current causes of women's underrepresentation in science. *Proc Natl Acad Sci USA*. **108**, 3157 (2011).

4. J. Huang, A. J. Gates, R. Sinatra, A.-L. Barabási, Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*. **117**, 4609–4616 (2020).

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

6. S. J. Ceci, D. K. Ginther, S. Kahn, W. M. Williams, Women in academic science: A changing landscape. *Psychological science in the public interest*. **15**, 75–141 (2014).

7. J. R. Cole, H. Zuckerman, The Productivity Puzzle: Persistence and Change in Patterns of Publication Among Men and Women Scientists In: Steinkamp, MW, Maehr, M.(Eds.): Advances in Motivation and Achievement (1984).

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

9. Y. Xie, K. A. Shauman, Sex differences in research productivity: New evidence about an old puzzle. *American sociological review*, 847–870 (1998).

10. L. A. Hechtman, N. P. Moore, C. E. Schulkey, A. C. Miklos, A. M. Calcagno, R. Aragon, J. H. Greenberg, NIH funding longevity by gender. *Proc Natl Acad Sci USA*. **115**, 7943 (2018).

11. T. Jr. Ley, B. H. Hamilton, The Gender Gap in NIH Grant Applications. *Science*. **322**, 1472–1474 (2008).

12. M. W. Nielsen, Limits to meritocracy? Gender in academic recruitment and promotion processes. *Science and Public Policy*. **43**, 386–399 (2016).

13. S. Chai, A. Menon, Breakthrough recognition: Bias against novelty and competition for attention. *Research Policy*. **48**, 733–747 (2019).

14. J. Wang, R. Veugelers, P. Stephan, Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy* (2017).

15. D. W. Aksnes, K. Rorstad, F. Piro, G. Sivertsen, Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of the American Society for Information Science and Technology*. **62**, 628–636 (2011).

16. S. R. Hutson, Gendered Citation Practices in American Antiquity and Other Archaeology Journals. *American Antiquity*. **67**, 331–342 (2002).

17. J. S. Long, Measures of Sex Differences in Scientific Productivity*. *Social Forces*. **71**, 159–178 (1992).

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

19. J. D. Dworkin, K. A. Linn, E. G. Teich, P. Zurn, R. T. Shinohara, D. S. Bassett, The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*. **23**, 918–926 (2020).

20. N. Caplar, S. Tacchella, S. Birrer, Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy*. **1**, 0141 (2017).

21. K. Wang, Z. Shen, C. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, R. Rogahn, A review of microsoft academic services for science of science studies. *Frontiers in Big Data*. **2**, 45 (2019).

22. M. M. King, C. T. Bergstrom, S. J. Correll, J. Jacquet, J. D. West, Men Set Their Own Cites High: Gender and Self-citation across Fields and over Time. *Socius*. **3**, 2378023117738903 (2017).

23. S. Mishra, B. D. Fegley, J. Diesner, V. I. Torvik, Self-citation is the hallmark of productive authors, of any gender. *PLOS ONE*. **13**, e0195773 (2018).

24. P. Azoulay, F. B. Lynn, Self-citation, cumulative advantage, and gender inequality in science. *Sociological science*. **7**, 152–186 (2020).

25. Z. Shen, H. Ma, K. Wang, A web-scale system for scientific knowledge exploration. *arXiv preprint arXiv:1805.12216* (2018).

26. L. Holman, C. Morandin, Researchers collaborate with same-gendered colleagues more often than expected across the life sciences. *PLoS One*. **14**, e0216128–e0216128 (2019).

27. E. Lee, F. Karimi, C. Wagner, H.-H. Jo, M. Strohmaier, M. Galesic, Homophily and minority-group size explain perception biases in social networks. *Nature Human Behaviour*. **3**, 1078–1087 (2019).

28. I. Newton, Letter to Robert Hooke (1676).

29. M. L. Weitzman, Recombinant Growth. *The Quarterly Journal of Economics*. **113**, 331–360 (1998).

30. L. Fleming, Recombinant uncertainty in technological search. *Management Science*. **47**, 117–132 (2001).

Supplementary Materials 1 – Dataset and Variable Construction

To explore gender difference in forward citations, we combined the PubMed[6] and Microsoft Academic Graph (MAG) dataset, using linkage available in MAG. PubMed is a bibliographic database for publications in the life sciences maintained by the National Library of Medicine that covers a full set of fields ranging from biostatistics and psychology to biomedicine and clinical sciences. It provides information about publications including article title, journal title, publication year, author names, author position, publication language, and publication type. A common issue, however, with bibliographic databases is that it is difficult to tell whether the "John Smith" listed on two articles are the same individual. We therefore make use of disambiguated names provided by MAG (*1*) to track the same author across multiple articles. This allows us to map authors' career history such as the year of first publication to compute experience, identify self-citations and track coauthor networks. To calculate an article's forward citations and its journal impact factor[7], we make use of the citing-to-cited article linkage from MAG. Until 2001, PubMed reported the last name and up to two first initials for each author. Starting from 2002, full first names were recorded whenever available[8]. We therefore restrict our focal sample to journal articles published between 2002 and 2017. Since our data ends in November 2021, publications from 2018 to 2020 are used to compute three-year forward citations for the 2017 focal articles.

For this period of interest, about 95% of articles are matched one-to-one between MAG and PubMed. We therefore keep these articles in our sample and drop the remaining non-one-to-one matches. We further restrict our sample to English publications (93.9% of PubMed articles) to avoid cross-language barriers in knowledge spillover. Moreover, we only include journal articles and exclude reviews, books, conference papers, letters, and editorial materials from our sample as their citation patterns may exhibit different trends. We also restrict our sample to the set of articles with no more than 17 authors (the 99^{th} percentile of the distribution of the number of authors) to rule out the concern that both the nature of research and the effects of authors' gender composition on forward citations may be different for very large teams. Articles with missing data were dropped. We further restrict our sample of focal articles, i.e., cited articles, to those with US-based first authors to focus on the biggest community of life scientists (28.8% of PubMed articles) and avoid issues that may stem from the significant country variances in the share of women's participation in the life sciences (*2*). We, however, do not restrict the country affiliation of the citing articles. It also enables us to increase the percentage of articles for which we can impute gender. For instance, we can impute both first and last author gender for 87% of this US-based sample, but only 79% for non-US based articles[9]. This is mainly due to the non-US based sample's higher proportion of articles with notoriously hard to genderize first names such as those of East Asian descent. Our results remain robust to enlarging our sample to OECD countries minus South Korea and Japan[10]. We also restrict our sample to articles where we can identify the gender of both first and last authors. All together our final sample consists of 2,432,806 focal publications.

To infer the gender of authors, we relied on their first names and the Genderize.io API (*3*). For every inputted first name, the API outputs a probability (between 0.5 and 1) that the name is female, as

[6] We use the Pubmed API available herein https://www.nlm.nih.gov/databases/download/pubmed_medline.html to download the data in bulk

[7] Pubmed does not provide the citing-to-cited article linkage, except for when both the citing article and the cited article belong to Pubmed Central, a subset of Pubmed.

[8] Full first names are recorded as long as they appear at the author position in the article. If the full first names are not reported at the author position but appear somewhere in the middle of the article, Pubmed does not record it.

[9] Using imputation thresholds of frequency>=10.

[10] All robustness findings can be obtained from the corresponding author.

well as the number of times the name appears in its database. There is also the option to input geographical information, namely the country, along with the first name. With this, the API in turn provides a country-specific gender probability and frequency. This country information can be useful to distinguish between cross-language differences in gender, such as "Jean" and "Andrea." "Jean" in France is more likely to be a male name, while more likely a female name in the US. Similarly, "Andrea" in Italy is more likely to be a male name, while more likely a female name in the US. Although its usefulness is limited, as a French person with the name of Jean may have a U.S. affiliation, we still use country affiliation whenever possible.

We imputed the gender of authors as follows. We start with unique author-affiliation level data and extract the first name and country information. We then perform data cleansing on the name including handling running initials, suffixes, prefixes, spaces, hyphens, quotes, non-English letters, and multiple first names, etc. We then input each first name-country level observation into the Genderize.io API and do so twice – once with country information and once without. In the absence of country information, Genderize.io provides a "global" frequency of the name, the global gender and its probability. With country information, Genderize.io provides a country-specific frequency and the corresponding gender and probability. The imputed country-specific frequency is always smaller than the global frequency, and there are therefore more chances for the API to return a null result when a first name is country restricted. We use the threshold of frequency>=10 to infer gender for both imputations – global and country-specific. In cases where the global imputed gender of a first name conflicts with its country-specific gender, we use the latter. Finally, different genders may be imputed for the same individual because a unique author may have different names (due to spelling differences and whether first names are reported or not) or multiple affiliations (due to career moves or multiple appointments in different countries). In these cases, we use the modal gender to reconcile conflicts when aggregating imputed gender for each unique author.

We repeat all our analyses using two additional sets of thresholds, one stricter and one laxer, and find robust and stable results throughout. The stricter threshold uses frequency>=10 and probability >= 0.9, which yields first and last author gender imputations of 69% for US-based articles and 64% for non-US based articles. The laxer criterion does not impose restrictions on frequency nor probability, and yields first and last author gender imputations of 94% for US-based articles and 88% for non-US based articles.

Our analysis is performed at the article level. For our main outcome variables of interest, we use three-year *forward citations* to study the gender citation gap. Three-year citations is a good predictor of both five-year and ten-year citations – the correlation between three-year citations and five-year citations is 0.9771, and that between three-year and ten-year citations is 0.8884. We repeat all our analyses using five-year and ten-year citation windows and find robust results. We use different variations of this *forward citation* variable in our analyses, including:

- *overall forward citations*, the total number of citations an article garners,
- *self-citations*, the number of citations an article garners from the citations of any of its authors,
- *male forward citations*, the number of citations an article garners from male authors, and
- *female forward citations*, the number of citations an article garners from female authors.

The explanatory variables of interest pertain to the inferred gender of the article's authors. We, thus create, the following indicator variables:

- *first author female* equals 1 if the article's first author is a woman, and
- *last author female* equals 1 if the article's last author is a woman.

In analyses on gender homophily, we investigate the role of the collaborative network and also include an indicator on whether the article reports first name or not. We use the following explanatory variables:

- *number of collaborators*, the total number of collaborators authors of the focal publication have published with in the three years prior,
- *share of female collaborators*, the percentage of female collaborators authors of the focal publication have published with in the three years prior,
- *no name last author* equals 1 if the article does not report first names of authors, and
- *no name first author* equals 1 if the article does not report first names of authors.

For each article observation, we also extract the following control variables that pertain to both article and author characteristics, to help partial out potential spurious relationships identified in prior research between our outcome and explanatory variables:

- *author team size* measured as the number of co-authors in the article,
- *number of references* the article makes,
- *number of pages* the article runs,
- *journal impact factor* in which the article is published, calculated for year t as the number of times articles from years t-1 and t-2 were cited by other articles during year t, divided by the number of articles published during years t-1 and t-2,
- *first author's experience* measured in years since first publication,
- *first author's cumulative publications* since first publication,
- *last author's experience* measured in years since first publication, and
- *last author's cumulative publications* since first publication.

We also include the following fixed effects:

- publication *year*,
- broad *field*,
- refined *concept*, and
- *journal.*

Tables S1 and **S2** include summary statistics and differences by first and last author gender for the key variables in our analysis.

We also present herein details of the various data sources used in **Fig 1A**, as follows:
Both the Survey of Earned Doctorates (SED) and Survey of Doctorate Recipients (SDR) are confidential, so we rely on the listed websites below that make certain tabulations publicly available.

- The share of women among earned doctorates in life sciences by degree year (1966-2020) is from the Survey of Earned Doctorates. https://ncsesdata.nsf.gov/home
- The share of women among SHE doctorate holders in life sciences employed in academia as full-time junior faculty and that employed in academia as full-time senior faculty during 1973-2010 are from NSF Science and Engineering Indicators 2014, table 5-15 at http://www.nsf.gov/statistics/seind14/index.cfm/appendix/tables.htm#c3. (This is the last year for which this tabulation is available. Latter versions of NSF S&E indicators no longer tabulate this.) The tabulation source is described as follows: "National Science Foundation, National Center for Science and Engineering Statistics, special tabulations (2013) of the Survey of Doctorate Recipients (various years)."

  Academic employment is limited to U.S. doctorate holders employed at 2- or 4-year colleges or universities. Full-time senior faculty includes full professors and associate professors. Full-time junior faculty includes assistant professors and instructors from 1973 to 1995; from 1997 to 2010, full-time junior faculty includes assistant professors.
- The share of articles collected by PubMed with female first and last author uses the same sample construction for all analyses as elaborated earlier.

Table S1. Gender differences in variables of interest, classified by *last* author gender

	Last Author Gender			
	Male	Female	Gap	
Number of Articles	1,795,497	637,309		
	73.80%	26.20%		
Number of 3-Year Forward Citations	11.05	9.12	1.92	***
	(25.80)	(17.55)		
Article Characteristics				
Number of References	30.21	29.90	0.31	***
	(31.82)	(30.93)		
Number of Pages	8.12	8.55	-0.43	***
	(5.31)	(5.51)		
Journal Impact Factor	3.49	3.07	0.43	***
	(3.19)	(2.87)		
Author Characteristics				
Author Team Size	4.63	4.30	0.33	***
	(2.96)	(2.82)		
Last Author Experience	17.48	13.07	4.41	***
	(10.66)	(10.35)		
Last Author Cumulated Publications	69.84	37.37	32.47	***
	(94.01)	(55.43)		
Number of Female at Non-Last Position	1.12	1.44	-0.32	***
	(1.39)	(1.57)		
Share with Female First Author[+]	33.0%	49.5%	-16.5%	***
First Author Experience	6.97	7.69	-0.72	***
	(8.23)	(8.88)		
First Author Cumulated Publications	16.48	17.51	-1.03	***
	(41.69)	(42.88)		

Note: [+]We focus on articles with at least two authors, when we summarize first author characteristics by last author gender

Table S2. **Gender differences in variables of interest, classified by *first* author gender**

	First Author Gender			
	Male	Female	Gap	
Number of Articles	1,544,727	888,079		
	63.50%	36.50%		
Number of 3-Year Forward Citations	10.93	9.87	1.06	***
	(23.60)	(24.47)		
Article Characteristics				
Number of References	29.64	30.97	-1.33	***
	(32.12)	(30.63)		
Number of Pages	8.07	8.51	-0.44	***
	(5.42)	(5.26)		
Journal Impact Factor	3.47	3.23	0.24	***
	(3.22)	(2.92)		
Author Characteristics				
Author Team Size	4.51	4.60	-0.09	***
	(2.95)	(2.88)		
First Author Experience	8.86	6.28	2.58	***
	(9.70)	(7.63)		
First Author Cumulated Publications	24.44	11.68	12.76	***
	(56.20)	(27.35)		
Number of Female at Non-First Position	0.94	1.39	-0.44	***
	(1.29)	(1.54)		
Share with Female Last Author^{+}	20.7%	34.2%	-13.5%	***
Last Author Experience	16.40	16.74	-0.33	***
	(10.61)	(10.65)		
Last Author Cumulated Publications	63.62	62.22	1.40	***
	(88.43)	(85.66)		

Note: $^{+}$We focus on articles with at least two authors, when we summarize last author characteristics by first author gender.

Supplementary Materials 2 – Replication of Gender Gap Findings

Using the dataset constructed above, we first replicate prior findings documenting the existence of a gender gap in forward citations (*2*, *4*). **Figs 1B** and **1C** in the main manuscript show the forward citation gap by gender and year of publication for women-led vs. men-led articles. **Table S3** shows the citation distribution by last author gender. We find that women-led articles receive 1.93 fewer citations than men-led articles. This gap holds throughout the citation distribution and is bigger at higher percentiles – it is one citation at the median and 19 citations at the 99th percentile. Articles with women last author are also more likely to receive no citation at all three years after publication. In **Table S4** we compare forward citations by first author[11] gender. Similarly, although smaller than the last author gender citation gap, articles with female first author receive on average 1.06 fewer citations than articles with male first author. The forward citation gap by first author gender is 15 citations at the 99th percentile, three at the 90th, but zero at 75th percentiles, median, and at 25th percentile.

Table S3. Gender differences in three-year forward citations received, classified by *last* author gender

Last Author	Num. of Obs.	Number of Three-Year Forward Citations						Share with Zero Citation
		Mean	P99	P90	P75	P50	P25	
Male	1,795,497	11.05	87	25	13	6	2	11.07%
Female	637,309	9.12	68	21	11	5	2	13.11%

Table S4. Gender differences in three-year forward citations received, classified by *first* author gender

First Author	Num. of Obs.	Number of Three-Year Forward						Share with Zero Citation
		Mean	P99	P90	P75	P50	P25	
Male	1,544,727	10.93	87	25	12	5	2	11.79%
Female	888,079	9.87	72	22	12	5	2	11.29%

[11] Note that when a paper has only one author, the author is treated as both last author and first author.

Supplementary Materials 3 – Prior Sources to Explain the Gender Gap

What explains the gender gap in forward citations? We first examine how articles written by the two genders differ in a set of observable factors identified in prior research, including author experience, size and composition of author team, journal, number of pages, number of backward references and academic field.

Despite fascinating anecdotes of young geniuses making scientific breakthroughs (*5*), the number of citations an article receives generally increases with author experience (*6*). Experienced researchers not only have more know-how in doing research but are also among a highly selected group of individuals who have survived tough competition and are better known to potential citers. However, even though women have recently caught up with men in terms of earned doctorates, women are still less likely to make it through the academic career ladder (*2*). As a result, women are on average less experienced than men among active scientists. **Tables S1** and **S2** also illustrate this discrepancy, where women last authors have on average approximately 13.07 years of experience and 37.37 publications, whereas men last authors have 17.48 years of experience and 69.84 publications. Similarly, women first authors have on average approximately 6.28 years of experience and 11.68 publications, whereas men first authors have on average 8.86 years of experience and 24.44 publications.

Differences in the number of authors per article may also contribute to the gender gap in forward citations, as women are more likely to work alone or work in smaller teams (*7*) but articles with more authors tend to receive more citations (*6*, *8*). However, in our sample of life scientists, we find that women and men have similar willingness to collaborate, as articles with women last author have on average only 0.33 fewer coauthors than articles with men last author, while articles with women first authors have 0.09 more coauthors as shown in **Tables S1** and **S2**. Team composition is another factor that drives forward citations. Diversity in various dimensions such as ethnicity (*9*), affiliations (*10*), and research topics (*11*) are associated with better publishing performance in terms of journal impact factor and forward citations. Along the same logic, if researchers of the same gender tend to coauthor together, they will not reap the benefits of the diversification premium (*12*). Moreover, the disadvantage of any single woman researcher in getting citations can be magnified.

Women and men also publish in journals of different impact, which is a popular measure of the journal's quality and by association the article's quality. In our sample, articles by men have higher journal impact factor than those by women – 3.49 vs. 3.07 by last author gender classification and 3.47 vs. 3.23 by first author gender classification (also shown in **Tables S1** and **S2**). The relationship between where an article is published and the citations the article receives is two-fold. First, getting published in a high impact journal may better signal the article's scientific merit, although not every article published in top journals is highly cited while articles rejected by top journals sometimes turn out to be pathbreaking. Second, high impact journals have broader audiences and thus have enhanced visibility and promotion effects for the articles published in them. Other factors that may affect citation rates and explain the gender gap include the length of articles and reference lists. Longer articles may contain more knowledge therefore increasing the number of references, and longer reference lists may indicate more thorough research. We therefore also control for these (also in **Tables S1** and **S2**).

Women and men also tend to work in different fields. For instance, women account for 10.1% of last authors in "Optics" and 66.7% in "Nursing." **Table S5**'s left panel lists the top ten fields for both women and men[12]. A field is the first level subject category classification from MAG. There are in total

[12] We rank the subject categories with at least 15,000 journal articles by their women shares of last authors. The top ten subject categories are the top ten women's fields, whereas the last 10 subject categories are the top 10 men's fields.

288 distinct fields averaging 8,447 articles per field that are systematically categorized by librarians. MAG also assigns *concepts* that are much more refined using its own topic modelling algorithms. In fact, there are 59,411 distinct refined concepts in our sample averaging 41 articles per concept. Each article is assigned concepts at different levels and a weight score associated to each concept. We selected the concept with the highest score as the concept that best describes the article. If fields or concepts with a female majority inherently have less active publications, then we would observe that women receive fewer citations on average thus contributing to the gender citation gap (*13*). Therefore, controlling for subject category is crucial and we opt to control for the refined concepts in all our regressions for conservatism.

By aggregating these factors into OLS regressions, **Table S6** explores how they contribute to the citation gap by last author gender (**Table S7** replicates the analysis for first author gender). We therefore regress the number of forward citations a focal article receives on the indicator variable of *last author female* while adding control variables step-by-step in each model. For all regression models in the manuscript, we also used OLS regressions with the natural logarithm plus one transformation on the dependent variables, the inverse hyperbolic transformation as well as Poisson regressions which have all yielded qualitatively similar and robust results. Model 1 does not include any controls except for the constant term, and thus replicates the average gap of 1.93 citations shown in **Table S1**. Model 2 controls for *publication year* and *concept* dummies, shrinking the negative coefficient of female last author by 33.8%. This suggests that women tend to publish in fields where articles are generally less cited. Adding in the *number of authors*, the *number of references* and the *number of pages* in Model 3, the coefficient of interest shrinks slightly by 9.5%. In Model 4, we control for last author's experience using both the number of elapsed *years* and *cumulative publications* since the author's first publication and find that the negative coefficient on women last author shrinks by 30.1%. When controlled in separate regressions, coefficients for each of the two experience variables are positive and significant. When controlled together, the sign on last author experience in years turns insignificant and negative while the sign on cumulative publications remains positive, suggesting the importance of publication intensity over the career. Model 5 further controls for *first author's experience*[13]. We also controlled for the *journal impact factor* in Model 6. This step decreases the negative coefficient of women last author by 9.8%. Since journals of high impact factor have higher standards with regard what they publish but also have promotion effects to the articles published, it is open to debate whether impact factor should be controlled for or not. When we add *journal fixed effects* in Model 7, the coefficients are similar and only shrink by 3%. In Model 8, instead of including concept field effects we use the broad field fixed effects and find similar results.

Taken together, if two articles are published in the same journal and have the same author experience, the same author team size, the same number of pages, and the same length of reference list, the article with a man last author receives more citations than the same article with a woman last author thus exhibiting a gender bias in received forward citations. In sum, 86% of the forward citation gap by last author gender can be explained by difference in observables (76% if gender is classified by first author see **Table S7**), with 33.8% due to academic field and year of publication, 9.5% due to article characteristics, 30.1% due to author experience, and another 12.9% due to journal.

Another factor debated in the literature that may also contribute to the gender bias in forward citations is self-citation (*14–16*). Self-citations are often used to increase the publication's visibility to other potential citers and subsequently further increase overall total citations. We measure self-citations as forward citations made by any author in the author team of the focal paper. Since men publish more often, they have more opportunities to self-cite especially if they work in a narrow set of subject categories. In a

---

[13] The coefficient on *female last author* increases slightly by 0.005, which suggests that last authors of both genders are not biased in picking first authors of very different "ability" (in terms of obtaining citations).

three-year window after publication, a focal article with male last author receives on average 11.05 total citations, among which 1.74 are from citing articles that share at least one common author with the focal article; while an article with female last author receives on average 9.12 total citations, among which 1.47 are from citing articles that share at least one common author with the focal article. Male last author articles thus receive 0.36 more self-citations than female last author articles, as shown in **Table S8** and **Fig S1**. Similarly male first author articles receive 0.20 more self-citations than female first author articles, as shown in **Table S9** and **Fig S2**. These trends persist in regressions, where we regress the number of self-citations a focal article receives on the indicator variables of *last author female* or *first author female* while controlling for observables using OLS. Models 1 and 2 in **Table S10** show that women-led articles receive fewer self-citations than men-led articles when the gender of the focal article is classified using both last and first authors. Models 3 and 4 replicate these same regressions while controlling for the full set of observable factors.

Table S5. Top 10 men and women's fields in the life sciences (among field with >= 15,000 articles)

Top Men's Field

Rank	Field	Num. of Articles	Share with Female Last Author	Share of Citations from Other Female: Written by Male Last Author	Written by Female Last Author	Gap	P-Value
				(1)	(2)	(1)-(2)	
1	Optics	18,625	10.1%	13.0%	15.8%	-2.9%	0.00
				(0.24)	(0.26)		
2	Surgery	98,887	13.4%	14.7%	21.0%	-6.3%	0.00
				(0.24)	(0.28)		
3	Cardiology	33,799	13.5%	15.4%	21.9%	-6.5%	0.00
				(0.24)	(0.28)		
4	Stereochemistry	15,677	14.2%	17.7%	20.5%	-2.8%	0.00
				(0.25)	(0.27)		
5	Artificial intelligence	15,813	15.9%	18.0%	22.8%	-4.9%	0.00
				(0.25)	(0.28)		
6	Chemical engineering	23,273	17.2%	18.9%	21.1%	-2.2%	0.00
				(0.25)	(0.26)		
7	Biophysics	36,128	17.4%	19.9%	22.7%	-2.8%	0.00
				(0.22)	(0.24)		
8	Radiology	48,813	18.6%	18.4%	25.7%	-7.3%	0.00
				(0.26)	(0.29)		
9	Computational biology	36,776	18.9%	21.5%	24.8%	-3.3%	0.00
				(0.22)	(0.24)		
10	Chromatography	21,548	19.2%	22.0%	26.6%	-4.6%	0.00
				(0.27)	(0.29)		

Top Women's Field

Rank	Field	Num. of Articles	Share with Female Last Author	Share of Citations from Other Female			
				Written by Male Last Author	Written by Female Last Author	Gap	P-Value
				(1)	(2)	(1)-(2)	
1	Nursing	28,663	66.7%	52.4%	66.1%	-13.8%	0.00
				(0.35)	(0.35)		
2	Developmental psychology	25,395	53.2%	47.7%	53.9%	-6.1%	0.00
				(0.32)	(0.32)		
3	Gerontology	16,544	52.3%	45.5%	52.2%	-6.7%	0.00
				(0.31)	(0.32)		
4	Medical education	35,489	47.0%	39.2%	52.9%	-13.6%	0.00
				(0.35)	(0.37)		
5	Clinical psychology	56,911	45.4%	42.2%	48.8%	-6.6%	0.00
				(0.30)	(0.31)		
6	Family medicine	51,891	44.5%	41.4%	50.3%	-8.9%	0.00
				(0.32)	(0.33)		
7	Obstetrics	16,994	44.4%	39.6%	48.8%	-9.2%	0.00
				(0.31)	(0.32)		
8	Demography	24,821	42.7%	40.1%	46.7%	-6.6%	0.00
				(0.30)	(0.31)		
9	Environmental health	22,900	40.5%	37.4%	43.8%	-6.5%	0.00
				(0.29)	(0.31)		
10	Pediatrics	22,121	36.4%	34.4%	39.9%	-5.5%	0.00
				(0.30)	(0.31)		

Table S6. OLS regression of three-year forward citations received on female author and observables, classified by *last* author gender

Dep. Var.:	Number of Three-Year Forward Citations							
OLS Regressions	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female Last Author	-1.925***	-1.274***	-1.092***	-0.513***	-0.518***	-0.330***	-0.271***	-0.297***
	(0.071)	(0.044)	(0.045)	(0.042)	(0.042)	(0.033)	(0.037)	(0.065)
Author Team Size			0.964***	0.912***	0.949***	0.726***	0.649***	0.692***
			(0.012)	(0.013)	(0.013)	(0.023)	(0.012)	(0.039)
Number of References			0.144***	0.140***	0.139***	0.106***	0.112***	0.118***
			(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.010)
Number of Pages			0.018	0.027*	0.030**	0.049***	0.303***	0.294***
			(0.015)	(0.015)	(0.015)	(0.015)	(0.018)	(0.033)
Last Author Experience				0.002	-0.004*	-0.028***	-0.024***	-0.032***
				(0.003)	(0.003)	(0.004)	(0.002)	(0.006)
Last Author Cumulated Pubs				0.018***	0.015***	0.011***	0.010***	0.011***
				(0.001)	(0.001)	(0.000)	(0.000)	(0.002)
First Author Experience					0.047***	0.027***	0.016***	0.012***
					(0.003)	(0.003)	(0.003)	(0.003)
First Author Cumulated Pubs					0.019***	0.016***	0.013***	0.014***
					(0.001)	(0.001)	(0.001)	(0.002)
Journal Impact Factor						2.291***	0.721***	0.734***
						(0.172)	(0.144)	(0.125)
Year FE	N	Y	Y	Y	Y	Y	Y	Y
Field FE	N	N	N	N	N	N	N	Y
Concept FE	N	Y	Y	Y	Y	Y	Y	N
Journal FE	N	N	N	N	N	N	Y	Y
Constant	11.048***	10.877***	1.968***	0.956***	0.315***	-4.779***	-1.264**	-1.509***
	(0.097)	(0.012)	(0.097)	(0.096)	(0.099)	(0.437)	(0.496)	(0.565)
Observations	2,432,806	2,432,806	2,432,806	2,432,806	2,432,806	2,432,806	2,432,806	2,432,806
R-squared	0.001	0.095	0.135	0.138	0.141	0.207	0.259	0.221

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table S7. OLS regression of three-year forward citations received on female author and observables, classified by *first* author gender

Dep. Var.:	Number of Three-Year Forward Citations							
OLS Regressions	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female First Author	-1.059***	-0.998***	-1.135***	-0.659***	-0.659***	-0.390***	-0.259***	-0.277***
	(0.062)	(0.045)	(0.043)	(0.042)	(0.042)	(0.054)	(0.039)	(0.081)
Author Team Size			0.973***	1.003***	0.951***	0.727***	0.650***	0.693***
			(0.012)	(0.013)	(0.013)	(0.023)	(0.012)	(0.039)
Number of References			0.144***	0.141***	0.139***	0.106***	0.112***	0.118***
			(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.010)
Number of Pages			0.019	0.024*	0.031**	0.049***	0.303***	0.294***
			(0.014)	(0.015)	(0.015)	(0.015)	(0.018)	(0.033)
First Author Experience				0.036***	0.043***	0.025***	0.014***	0.010***
				(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
First Author Cumulated Pubs				0.025***	0.018***	0.015***	0.013***	0.014***
				(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Last Author Experience					-0.001	-0.026***	-0.023***	-0.031***
					(0.002)	(0.004)	(0.002)	(0.006)
Last Author Cumulated Pubs					0.016***	0.011***	0.010***	0.011***
					(0.001)	(0.000)	(0.000)	(0.002)
Journal Impact Factor						2.290***	0.721***	0.734***
						(0.172)	(0.144)	(0.126)
Year FE	N	Y	Y	Y	Y	Y	Y	Y
Field FE	N	N	N	N	N	N	N	Y
Concept FE	N	Y	Y	Y	Y	Y	Y	N
Journal FE	N	N	N	N	N	N	Y	Y
Constant	10.930***	10.908***	2.053***	0.994***	0.387***	-4.744***	-1.260**	-1.507***
	(0.093)	(0.016)	(0.108)	(0.111)	(0.110)	(0.458)	(0.512)	(0.576)
Observations	2,432,806	2,432,806	2,432,806	2,432,806	2,432,806	2,432,806	2,432,806	2,432,806
R-squared	0.000	0.095	0.135	0.138	0.141	0.207	0.259	0.221

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table S8. Gender differences in three-year self-citations received, classified by *last* author gender

	All	Last Author		
		Male	Female	Male-Female
Number of Self Citations	1.74	1.83	1.47	0.36 ***
	(2.88)	(3.00)	(2.51)	

Table S9. Gender differences in three-year self-citations received, classified by *first* author gender

	All	First Author		
		Male	Female	Male-Female
Number of Self Citations	1.74	1.81	1.61	0.20 ***
	(2.88)	(3.01)	(2.63)	

Figure S1. Three-year self-citations received by year, classified by *last* author

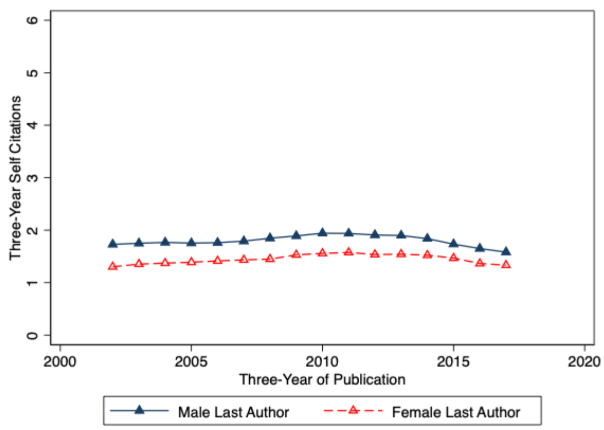


Figure S2. Three-year self-citations received by year, classified by *first* author

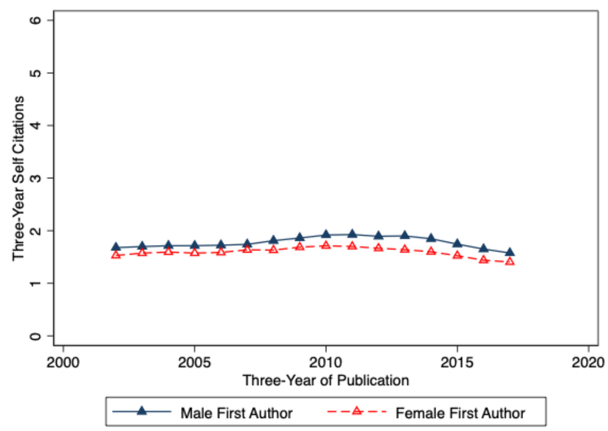


Table S10. OLS regression of three-year self-citations received by female authors, classified by *last* and *first* author gender

Dep. Var.:	Number of Three-Year Self Citations			
OLS Regressions	(1)	(2)	(3)	(4)
Last Author Female	-0.358***		-0.054***	
	(0.011)		(0.004)	
First Author Female		-0.205***		-0.118***
		(0.009)		(0.004)
Author Team Size			0.218***	0.218***
			(0.002)	(0.002)
Number of References			0.009***	0.009***
			(0.000)	(0.000)
Number of Pages			0.053***	0.053***
			(0.001)	(0.001)
First Author Experience			0.000	-0.001
			(0.000)	(0.000)
First Author Cumulated Pubs			0.003***	0.003***
			(0.000)	(0.000)
Last Author Experience			-0.003***	-0.003***
			(0.000)	(0.000)
Last Author Cumulated Pubs			0.002***	0.002***
			(0.000)	(0.000)
Journal Impact Factor			0.069***	0.069***
			(0.003)	(0.003)
Year FE	N	N	Y	Y
Concept FE	N	N	Y	Y
Journal FE	N	N	Y	Y
Constant	1.831***	1.812***	-0.335***	-0.308***
	(0.013)	(0.013)	(0.015)	(0.015)
Observations	2,432,806	2,432,806	2,432,806	2,432,806
R-squared	0.003	0.001	0.273	0.274

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Supplementary Materials 4 – Gender Homophily in Citations Patterns

After self-citations are excluded, we focus on understanding who is citing whom – more specifically, which gender do women receive more citations from and which gender do men receive more citations from. In fact, we split the remaining forward citations a focal article receives into citations from male citers and those from female citers as dependent variables. For citing articles as well, we use both first and last author gender for classification into women-led or men-led. Gender differences in forward citations from male and female citers are shown in **Tables S11** and **S12**. They exhibit gender homophily where men-led articles tend to receive more citations from men, while women-led articles tend to receive more citations from women. Regression analyses indicate the same evidence. For instance, when we focus on forward citations from women citers classified by last author gender in Model 2 of **Table S13**, articles by women receive 0.187 more citations than articles written by men controlling for all observables except for field or concept fixed effects. Conversely, Model 6 focuses on citations from male citers as dependent variable and shows that articles by women receive 0.368 fewer citations than articles written by men. **Table S14** replicates the analysis for first author gender.

One potential explanation of this gender homophily in forward citations may simply be that the two genders sort into different fields. In the extreme, if the two genders are segregated into different fields, then, of course, women will receive citations only from women and men only from men. However, several robustness analyses help us rule out this sorting mechanism. The right panel of **Table S5** is a first indication. Even in the top 10 men's fields, women-led articles still receive a higher share of citations from women than men-led articles. Moreover, while controlling for all observables and whether we use the broad category with 288 fields or the refined topics with 59,411 distinct concepts as fixed effects, the coefficients of interest in Models 3, 4, 7 and 8 of **Tables S13** and **S14** are in the same direction and at the same level of significance as the regressions without controls and fixed effects. These regressions correspond to **Fig 2B** that demonstrates the presence of gender homophily in forward citations, where men receive more citations from men and women receive more citations from women. This reassures us again that our results are not purely driven by sorting across refined research topics within a field.

Table S11. Gender differences in forward citations from male and female citers, classified by *last* author gender

	Last Author		
	Male	Female	Male-Female
Number of Articles	1,795,497	637,309	
	73.8%	26.2%	
Number of Forward Citations, Three-Year Window:			
All	11.05	9.12	1.92 ***
	(25.80)	(17.55)	
Non-Self	9.22	7.65	1.57 ***
	(24.59)	(16.42)	
From Female Last Author	2.06	2.26	-0.20 ***
	(5.60)	(5.14)	
From Male Last Author	6.45	4.86	1.59 ***
	(17.19)	(11.02)	
From Female First Author	3.11	3.16	-0.04 ***
	(8.40)	(6.98)	
From Male First Author	5.26	3.85	1.40 ***
	(14.05)	(8.82)	

Table S12. Gender differences in forward citations from male and female citers, classified by *first* author gender

	First Author		
	Male	Female	Male-Female
Number of Articles	1,544,727	888,079	
	63.5%	36.5%	
Number of Forward Citations, Three-Year Window:			
All	10.93	9.87	1.06 ***
	(23.60)	(24.47)	
Non-Self	9.12	8.26	0.85 ***
	(22.24)	(23.59)	
From Female Last Author	2.00	2.32	-0.32 ***
	(5.02)	(6.2)	
From Male Last Author	6.42	5.36	1.06 ***
	(16.00)	(15.49)	
From Female First Author	3.02	3.31	-0.30 ***
	(7.58)	(8.8)	
From Male First Author	5.26	4.24	1.02 ***
	(13.05)	(12.6)	

Table S13. OLS regression of three-year forward citations received from male and female citers on female author, classified by *last* author gender

Dep. Var.:	Three-Year Forward Citations from **Female Last** Author			
OLS Regressions	(1)	(2)	(3)	(4)
Female Last Author	0.201***	0.187***	0.157***	0.110***
	(0.017)	(0.011)	(0.011)	(0.010)
Observations	2,432,806	2,432,806	2,432,806	2,432,806
R-squared	0.000	0.129	0.165	0.208
Dep. Var.:	Three-Year Forward Citations from **Male Last** Author			
OLS Regressions	(5)	(6)	(7)	(8)
Female Last Author	-1.592***	-0.368***	-0.362***	-0.296***
	(0.046)	(0.028)	(0.025)	(0.024)
Observations	2,432,806	2,432,806	2,432,806	2,432,806
R-squared	0.002	0.150	0.192	0.249
Other controls	N	Y	Y	Y
Year FE	N	Y	Y	Y
Journal FE	N	Y	Y	Y
Field FE	N	N	Y	N
Concept FE	N	N	N	Y

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table S14. OLS regression of three-year forward citations received from male and female citers on female author, classified by *first* author gender

Dep. Var.:	Three-Year Forward Citations from **Female First** Author			
OLS Regressions	(1)	(2)	(3)	(4)
Female First Author	0.298***	0.210***	0.174***	0.132***
	(0.021)	(0.014)	(0.015)	(0.015)
Observations	2,432,806	2,432,806	2,432,806	2,432,806
R-squared	0.000	0.142	0.180	0.226

Dep. Var.:	Three-Year Forward Citations from **Male First** Author			
OLS Regressions	(5)	(6)	(7)	(8)
Female First Author	-1.024***	-0.321***	-0.302***	-0.248***
	(0.032)	(0.019)	(0.021)	(0.019)
Observations	2,432,806	2,432,806	2,432,806	2,432,806
R-squared	0.001	0.145	0.187	0.244
Other controls	N	Y	Y	Y
Year FE	N	Y	Y	Y
Journal FE	N	Y	Y	Y
Field FE	N	N	Y	N
Concept FE	N	N	N	Y

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Supplementary Materials 5 – Gender Homophily in Citations from Collaborator Networks

We delve further into possible reasons behind the observed gender homophily in forward citations. One possible reason stems from positive network externalities from within authors' professional networks. Indeed, close professional acquaintances – both informal relationships or formal collaborators– know each other's work better and are therefore more likely citers. Researchers may also cite a collaborator or friend's work even when only remotely related as a favor.

We use collaborator networks as proxy for professional networks and construct it by including researchers who have collaborated with one or more authors of a focal article's author team in the three years prior to the article's publication. In effect, we take articles published from 2005 to 2017, and compute degree-one collaborators of each article based on its authors' three prior years of publications and use the post publication three-year forward citations window as dependent variable. Thus, for articles published in 2005 for example, the collaborator network is computed using publications from 2002-2004, while forward citations are computed using citations from 2006-2008. **Tables S15** and **S16** show that articles by men have bigger collaborator networks than articles by women, irrespective of whether the article gender is classified by first or last author. The gender composition of collaborators also presents signs of gender homophily, where both genders have a higher share of same-gender collaborators than predicted by randomness, as shown in **Fig 3A**.

Table S17 examines how the gender differences in collaborator network affect forward citations made by the two genders based on last author gender. Model 5 shows that as the number of collaborators increases so do citations from female citers after controlling for all observables. Moreover, citations from women also increase significantly as the share of women among collaborators increases. Model 6 also shows that as the number of collaborators increase so do citations from male citers, however, citations from men decrease significantly as the share of women among collaborators increases. Models 5 and 6 and corresponds to the regression models shown in **Fig 3B**. **Table S18** replicates the analysis for first author gender.

The second possible reason behind the observed gender homophily in forward citations stem from gender biases outside authors' immediate collaborative networks. In **Table S19**, we show the regression results that correspond to **Fig 3C** for articles classified by either last or first author. Our sample in these regressions is a subset of articles that sometimes report author first names and sometimes do not, but whose gender classification we have imputed from our disambiguated author database. After controlling for observables and author fixed effects to ensure that we have the exact the same author, we find in Models 2 & 5 that citations from female citers to female articles not reporting first names are not significantly different from zero, and in Models 3 & 6 that citations from male citers to female articles not reporting first names are not significantly different from zero. This entails that for a same author, the reporting of first names or not (i.e. whether the gender could be guessed by strangers or not) does not change the degree of bias this author is subject to from both male and female citers. Thus, the source of the gender homophily does not appear to be from outside authors' immediate collaboration networks but from within.

Table S15. Gender differences in first-degree collaborative network, classified by last author

	Last Author Gender		Gap	
	Male	Female		
Composition of the Degree-one collaborators of author team (Three-Year Window, 2005-2017)				
Total	141.65	117.94	23.71	***
	(218.71)	(215.02)		
Number of Female	43.98	43.91	0.06	(P-Val: 0.65)
	(74.25)	(81.66)		
Number of Male	86.37	65.74	20.64	***
	(131.20)	(122.03)		
Number of Gender Unidentified	11.30	8.29	3.01	***
	(25.51)	(20.54)		
Female Share*	32.4%	42.6%	-10.3%	***

*female share = number of female/(number of female + male).

Table S16. Gender differences in first-degree collaborative network, classified by first author

	First Author Gender		Gap	
	Male	Female		
Composition of the Degree-one collaborators of author team (Three-Year Window, 2005-2017)				
Total	134.78	136.22	-1.44	***
	(214.02)	(224.48)		
Number of Female	41.07	48.82	-7.74	***
	(71.74)	(83.20)		
Number of Male	82.84	77.52	5.32	***
	(129.24)	(128.88)		
Number of Gender Unidentified	10.87	9.88	0.98	***
	(25.00)	(23.10)		
Female Share*	31.8%	40.6%	-8%	***

*female share = number of female/(number of female + male).

Table S17. OLS regression of three-year forward citations received from male and female citers (classified by *last* author gender) on female author classified by *last* author gender, the number of first-degree collaborators in one's three-year prior collaborative network and share of female among collaborators

Dep. Var.:	Number of Citations			Number of Citations		
	Non-Self	By Female Last Author	By Male Last Author	Non-Self	By Female Last Author	By Male Last Author
	(1)	(2)	(3)	(4)	(5)	(6)
Last Author Female	-0.216***	0.116***	-0.297***	-0.252***	0.080***	-0.298***
	(0.041)	(0.011)	(0.027)	(0.043)	(0.012)	(0.029)
Number of Degree-1 Collaborators (in hundreds)				0.430***	0.111***	0.286***
				(0.025)	(0.006)	(0.017)
Share of female among collaborators				0.418***	0.752***	-0.276***
				(0.136)	(0.036)	(0.090)
Control Variables	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Concept FE	Y	Y	Y	Y	Y	Y
Observations	1,944,513	1,944,513	1,944,513	1,944,513	1,944,513	1,944,513
R-squared	0.231	0.204	0.242	0.231	0.205	0.243

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table S18. OLS regression of three-year forward citations received from male and female citers (classified by *first* author gender) on female author classified by *first* author gender, the number of first-degree collaborators in one's three-year prior collaborative network and share of female among collaborators

Dep. Var.:	Number of Citations			Number of Citations		
	Non-Self	By Female First Author	By Male First Author	Non-Self	By Female First Author	By Male First Author
	(1)	(2)	(3)	(4)	(5)	(6)
First Author Female	-0.266***	0.169***	-0.397***	-0.149***	0.106***	-0.232***
	(0.065)	(0.024)	(0.034)	(0.044)	(0.017)	(0.022)
Number of Degree-1 Collaborators (in hundreds)				0.429***	0.165***	0.225***
				(0.025)	(0.009)	(0.014)
Share of female among collaborators				0.373***	0.878***	-0.457***
				(0.119)	(0.045)	(0.067)
Control Variables	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Concept FE	Y	Y	Y	Y	Y	Y
Observations	1,944,513	1,944,513	1,944,513	1,944,513	1,944,513	1,944,513
R-squared	0.176	0.172	0.179	0.231	0.222	0.238

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table S19. OLS regression of three-year forward citations received from male and female citers (classified by last and first author gender) on female author classified by *last* and *first* author gender interacted with no name *first* and *last* author

Dep. Var.:	Number of Three-Year Forward Citations					
	Non-Self	By Female Last Author	By Male Last Author	Non-Self	By Female First Author	By Male First Author
Sample:	Articles with ***last*** author who occasionally don't report first names			Articles with ***first*** author who occasionally don't report first names		
	(1)	(2)	(3)	(4)	(5)	(6)
No Name Last Author	-0.167	-0.007	-0.173*			
	(0.143)	(0.034)	(0.102)			
Last Author Female x No Name Last Author	-0.128	-0.092*	-0.019			
	(0.162)	(0.051)	(0.108)			
No Name First Author				-0.160	-0.023	-0.125**
				(0.102)	(0.039)	(0.062)
First Author Female x No Name First Author				0.132	0.021	0.093
				(0.123)	(0.052)	(0.069)
Control Variables	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Concept FE	Y	Y	Y	Y	Y	Y
Last Author FE	Y	Y	Y	N	N	N
First Author FE	N	N	N	Y	Y	Y
Observations	797,382	797,382	797,382	381,957	381,957	381,957
R-squared	0.305	0.297	0.317	0.475	0.438	0.483

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Supplementary Materials 6 – Robustness Analyses in Classification of Article Gender

Instead of gendering articles using women-led classifications by last and first author, we also classify an article's gender using several specifications to proxy the share of female authors in an article. Thus, on top of the *female first author* and *female last author* indicator explanatory variables, we also include the following:

- *number of female authors in non-last positions*,
- *number of female authors in middle positions*,
- *number of female authors in non-first positions*,
- *all male*, indicator variable equals 1 when all authors in the team are male,
- *minority female*, indicator variable equals 1 when female authors account for less than half of the authors in the team,
- *majority female*, indicator variable equals 1 when female authors account for half or more of the authors in the team,
- *all female*, indicator variable equals 1 when all authors in the team are female, and
- *share of female*, percentage of females in author team.

As shown in Tables **S20** and **S21**, our main finding of gender homophily in forward citations is robust to various gender classifications in the share of female authors in an article while controlling for observables. In these regressions, we still classify the gender of citers using women-led and men-led articles. All odd model specifications are citations made by women, while all even model specifications are citations made by men. The coefficients on all explanatory variables that proxy the share of female authors in an article are positive and significant on citations from women as illustrated in odd model specifications. This shows that women tend to receive citations from women more. Conversely, the same coefficients of interest are all negative and significant on citations from men in even model specifications, which shows that men tend to receive citations from men more.

We further decompose the forward citations received excluding self-citations into the:

- *number of citations from all male* articles,
- *number of citations from minority female* articles,
- *number of citations from majority female* articles, and
- *number of citations from all female* articles.

As shown in Tables **S22**, our main finding of gender homophily in forward citations is again robust this time to various gender classifications for citers of an article. While controlling for observables, Model 1 shows that compared to *all male* author teams, those in *minority female*, *majority female* or *all female* teams receive significantly less citations from *all male* articles. Thus, showing again that men receive more citations from men. Conversely, Models 3 and 4 show that compared to *all male* teams, those in *minority female*, *majority female* or *all female* team receive significantly more citations from *all female* articles or *majority female* articles. Thus, showing here that women tend to receive more citations from women.

Table S20.

Dep. Var.	Number of Three-Year Forward Citations, by Gender of ***Last*** Author							
	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Last Author Female	0.108***	-0.258***	0.095***	-0.277***				
	(0.011)	(0.028)	(0.012)	(0.030)				
N Female Authors in Non-Last Positions	0.071***	-0.217***						
	(0.006)	(0.015)						
First Author Female			0.113***	-0.146***				
			(0.013)	(0.030)				
N Female Authors in Middle Positions			0.063***	-0.175***				
			(0.005)	(0.014)				
Minority Female					0.046***	-0.053*		
					(0.011)	(0.032)		
Majority Female					0.198***	-0.337***		
					(0.018)	(0.042)		
All Female					0.277***	-0.251***		
					(0.017)	(0.037)		
Share of Female							0.339***	-0.521***
							(0.023)	(0.053)
Constant	-0.027	0.005	0.024	0.315	-0.116	0.115	-0.128	0.156
	(0.121)	(0.312)	(0.130)	(0.332)	(0.127)	(0.328)	(0.128)	(0.330)
Control Variables	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Concept FE	Y	Y	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y	Y	Y
Observations	1,937,093	1,937,093	1,777,242	1,777,242	1,937,093	1,937,093	1,937,093	1,937,093
R-squared	0.202	0.241	0.208	0.248	0.202	0.241	0.202	0.241

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table S21.

Dep. Var.	Number of Three-Year Forward Citations, by Gender of ***First*** Author							
	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First Author Female	0.135***	-0.196***	0.122***	-0.156***				
	(0.017)	(0.023)	(0.018)	(0.024)				
N Female Authors in Non-First Positions	0.057***	-0.223***						
	(0.006)	(0.010)						
Last Author Female			0.079***	-0.260***				
			(0.017)	(0.024)				
N Female Authors in Middle Positions			0.063***	-0.175***				
			(0.007)	(0.012)				
Minority Female					0.059***	-0.069***		
					(0.016)	(0.026)		
Majority Female					0.214***	-0.354***		
					(0.025)	(0.035)		
All Female					0.311***	-0.289***		
					(0.023)	(0.030)		
Share of Female							0.364***	-0.548***
							(0.032)	(0.044)
Constant	-0.085	0.014	0.025	0.294	-0.169	0.161	-0.178	0.200
	(0.185)	(0.266)	(0.191)	(0.272)	(0.186)	(0.269)	(0.188)	(0.271)
Control Variables	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Concept FE	Y	Y	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y	Y	Y
Observations	1,937,093	1,937,093	1,777,242	1,777,242	1,937,093	1,937,093	1,937,093	1,937,093
R-squared	0.219	0.237	0.224	0.244	0.219	0.237	0.219	0.237

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table S22.

Dep. Var.	Number of Three-Year Forward Citations			
	From All Male	From Minority Female	From Majority Female	From All Female
	(1)	(2)	(3)	(4)
Minority Female	-0.077***	-0.002	0.053***	0.019***
	(0.012)	(0.019)	(0.012)	(0.003)
Majority Female	-0.261***	-0.154***	0.177***	0.099***
	(0.013)	(0.026)	(0.018)	(0.005)
All Female	-0.230***	-0.090***	0.186***	0.160***
	(0.013)	(0.021)	(0.017)	(0.007)
Constant	0.383***	-0.285	-0.102	0.004
	(0.073)	(0.233)	(0.142)	(0.017)
Control Variables	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Concept FE	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y
Observations	1,937,093	1,937,093	1,937,093	1,937,093
R-squared	0.270	0.202	0.221	0.175

Robust standard errors clustered at concept level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

References

1. K. Wang, Z. Shen, C. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, R. Rogahn, A review of microsoft academic services for science of science studies. *Frontiers in Big Data*. **2**, 45 (2019).

2. J. Huang, A. J. Gates, R. Sinatra, A.-L. Barabási, Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*. **117**, 4609–4616 (2020).

3. L. Santamaría, H. Mihaljević, Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*. **4**, e156 (2018).

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

5. D. K. Simonton, Age and outstanding achievement: What do we know after a century of research? *Psychological Bulletin*. **104**, 251 (1988).

6. B. F. Jones, The burden of knowledge and the "death of the Renaissance Man": Is innovation hetting harder? *Review of Economic Studies*. **76**, 283–317 (2009).

7. A. Boschini, A. Sjögren, Is team formation gender neutral? Evidence from coauthorship patterns. *Journal of Labor Economics*. **25**, 325–365 (2007).

8. S. Wuchty, B. F. Jones, B. Uzzi, The Increasing Dominance of Teams in Production of Knowledge. *Science*. **316**, 1036–1039 (2007).

9. R. B. Freeman, W. Huang, Collaboration: Strength in diversity. *Nature*. **513**, 305 (2014).

10. B. F. Jones, S. Wuchty, B. Uzzi, Multi-university research teams: Shifting impact, geography, and stratification in science. *science*. **322**, 1259–1262 (2008).

11. S. Chai, A. Menon, Breakthrough recognition: Bias against novelty and competition for attention. *Research Policy*. **48**, 733–747 (2019).

12. M. W. Nielsen, S. Alegria, L. Börjeson, H. Etzkowitz, H. J. Falk-Krzesinski, A. Joshi, E. Leahey, L. Smith-Doerr, A. W. Woolley, L. Schiebinger, Opinion: Gender diversity leads to better science. *Proceedings of the National Academy of Sciences*. **114**, 1740–1742 (2017).

13. Kozlowski Diego, Larivière Vincent, Sugimoto Cassidy R., Monroe-White Thema, Intersectional inequalities in science. *Proceedings of the National Academy of Sciences*. **119**, e2113067119 (2022).

14. M. M. King, C. T. Bergstrom, S. J. Correll, J. Jacquet, J. D. West, Men Set Their Own Cites High: Gender and Self-citation across Fields and over Time. *Socius*. **3**, 2378023117738903 (2017).

15. S. Mishra, B. D. Fegley, J. Diesner, V. I. Torvik, Self-citation is the hallmark of productive authors, of any gender. *PLOS ONE*. **13**, e0195773 (2018).

16. P. Azoulay, F. B. Lynn, Self-citation, cumulative advantage, and gender inequality in science. *Sociological science*. **7**, 152–186 (2020).