Quantitative evaluation of gender bias in astronomical publications from citation counts

Neven Caplar*, Sandro Tacchella* and Simon Birrer

Numerous studies across different research fields have shown that both male and female referees consistently give higher scores to work done by men than to identical work done by women1–3. In addition, women are under-represented in prestigious publications and authorship positions4,5 and women receive ~10% fewer citations6,7. In astronomy, similar biases have been measured in conference participation8,9 and success rates for telescope proposals10,11. Even though the number of doctorate degrees awarded to women is constantly increasing, women still tend to be under-represented in faculty positions12. Spurred by these findings, we measure the role of gender in the number of citations that papers receive in astronomy. To account for the fact that the properties of papers written by men and women differ intrinsically, we use a random forest algorithm to control for the non-gender-specific properties of these papers. Here we show that papers authored by women receive 10.4 ± 0.9% fewer citations than would be expected if the papers with the same non-gender-specific properties were written by men.

We consider a complete sample of >200,000 publications from 1950 to 2015 from five major astronomy journals: Astronomy & Astrophysics, Astrophysical Journal, Monthly Notices of the Royal Astronomical Society, Nature and Science. We used the Smithsonian Astrophysical Observatory (SAO)/National Aeronautics and Space Administration (NASA) Astrophysics Data System (ADS) and the arXiv database to gather the following information about these papers: the names and number of authors; the number of references; the year of publication; the journal of publication; the abstract; and the name of the first author’s institution. We determine the gender of the first author by matching their name to a number of different publicly available databases. We clean the sample by removing entries for which we are not able to determine the gender of the first author. We also remove entries without references or citations. Our final dataset contains 149,741 papers. Further details about this procedure are available in the Methods.

Throughout the study we assume that men and women should receive the same number of citations for papers that have the same non-gender-specific properties. Any difference in the citation counts between papers led by men or women with matched non-gender properties is labelled as ‘gender bias’. For all practical purposes, the phrases ‘women’ and ‘men’ are to be understood as ‘first authors that we deduced to be women in this analysis’ and ‘first authors that we deduced to be men in this analysis’, respectively. Gender identities outside the male/female binary are not considered in the analysis.

We first examine whether there is a difference between men and women in terms of the number of citations. Papers written by men and women have different properties in the sample (see Methods). As the citation count is expected to correlate with certain non-gender-specific properties of the papers (such as seniority or number of references), we have to be careful when interpreting the quoted difference in the number of citations. We attempt to separate the gender bias effect from the effect caused by non-gender-specific properties of the papers.

Figure 1 shows the mean number of citations received by men divided by the mean number of citations received by women in a given year. We see a large difference between men and women in the early years of this study, with men receiving between 50 and 100% more citations than women. In this early period, the errors are large due to the small number of papers in total and even smaller numbers of papers authored by women. Overall, the difference has been decreasing over time. We also show the results of fitting the data with the functional form of $a_2 e^{t/\gamma} + b_h$, where $y$ is the year. The best-fit parameters are $a_1 = 0.06 \pm 0.02$, $a_2 = 0.38 \pm 0.24$, $a_3 = 1.00 \pm 0.04$ and $\gamma = 1974 \pm 12$. When written in this form, $a_2$ can be interpreted as the value of gender difference in the far future when the first term of the equation becomes negligible.

To quantify the difference in a single number, we introduce the variable, $b_h$, defined as a constant fit to the data presented in Fig. 1 after a certain year. In this work, we use 1985 as the cutoff year—that is, $b_h$ is obtained by fitting the data with a constant from 1985 to 2015. Thus we search for the value of $b_h$ that minimizes

$$\sum_{y=\text{year} \geq 1985} \left( \frac{d_y - b_{ym}}{\sigma_{d_y}} \right)^2$$

where $d_y$ is the gender difference measured in a given year ($y$) and $\sigma_{d_y}$ is the estimated error of the measured gender difference. Using this definition we find $b_{ym} = 1.056 \pm 0.010$. This means that men received around 6% more citations on average than women. Changing the cutoff year does not significantly change our results because the fit is always dominated by the data points in the latter years as a result of their smaller errors. For example, when taking the cutoff year to be 2000, we find $b_{ym} = 1.046 \pm 0.009$.

It is complex to estimate the amount of gender bias given the difference in the properties of papers written by men and women. Any difference that we see could just be a consequence of the fact that papers authored by men and women in the sample inherently differ in their properties and hence may receive fewer citations, not because of the authors’ gender, but because of some other parameter. Given that there are many plausible variables influencing the citation number of papers, it is impossible to isolate or study a single variable (such as seniority or number of references) to capture the full span of possibilities influencing our estimate of gender bias. Therefore, we resort to machine learning techniques to correct and estimate more accurately the amount of gender bias.

The main idea is to train the random forest algorithm13 on the sample of papers authored by men using all the non-gender-specific...
parameters available for the dataset. These non-gender-specific parameters include the seniority of the first author, the number of references, the total number of authors, the year of publication, the journal of publication, the field of study and the geographical region of the first author’s institution. With the predictor trained on the sample of papers written by men, we then estimate the number of expected citations for the papers written by women given the properties of their papers. By comparing the predicted number of citations with the measured number of citations, we are able to constrain the intrinsic gender bias, which is corrected for the non-gender-specific properties of papers.

Figure 2 shows the ratio of the measured number of citations that women have received to the number of citations that would be expected from our analysis. We find that papers written by women systematically receive fewer citations than would be expected given the other, non-gender-specific, properties of their papers. We also show the results of fitting the data with the functional form of \( b_0 e^{b_1 (y - y_0)} + b_2 \), where \( y \) is the year. The best-fit parameters are \( b_0 = 0.06 \pm 0.01, b_1 = 3 \pm 2, b_2 = 0.94 \pm 0.03 \) and \( y_0 = 1939 \pm 25 \). We define the quantity \( b_{ww} \), characterizing this difference between the simulated sample of papers written by women (w) and the actual sample of papers written by women (w). We measured \( b_{ww} \) fitting the data presented in Fig. 2 from 1985 with the same procedure as presented earlier. We find \( b_{ww} = 0.896 \pm 0.009 \)—that is, women systematically receive 10.4 \( \pm \) 0.9\% fewer citations than would be expected given the properties of their papers.

To check the consistency of the results presented here (bias that amounts to 10\%) and the uncorrected gender difference that amounts to 6\%, we replace the measured number of citations received for papers authored by women with the predicted number of citations. With this experiment, we measure what would be the difference in number of citations if there was no gender bias between papers written by men (m) and women (w). We measure this value to be \( b_{ww} = 0.958 \pm 0.008 \).

Figure 1 | Ratio of mean number of citations for papers written by men to the mean number of citations for papers written by women. The error bars are obtained by bootstrapping and denote the 1\% errors on the mean of the measurement. The shaded green area shows the 1\% uncertainty on the best fit (see text for details). The dashed line indicates unity. In every year since 1960, men receive on average similar or larger number of citations than women. This difference is higher at earlier times, although the measurement is also more uncertain. Since 1985, this difference is roughly consistent at ~6\%.

Methods

Data sources. To obtain a list of all published papers in the field of astronomy, we downloaded from the SAO/NASA ADS (http://adswww.harvard.edu/) all the entries available in the database ‘astronomy’ published between 1950 and 2015 in one of the following five journals: Astronomy & Astrophysics, Astrophysical Journal, Monthly Notices of the Royal Astronomical Society, Nature and Science. We choose these five journals because they encompass the vast majority of astronomical research today and are well-established journals with long historical records. The SAO/NASA astronomy API service provides many types of metrics for each paper. Specifically, we chose to download the names of the authors and their institutions, the number of citations, the number of references, the name of the publishing journal, the abstract of the paper and the year of publication. All information was downloaded in a single effort in June 2016 and therefore the number of citations for every paper reflects the state of the metric at that point in time.

We augmented the data with information available from the arXiv database (https://arxiv.org/) for papers where such data exist. For each paper found in the arXiv database, we recorded the designated field (‘astrophysics of galaxies,’ ‘cosmology and nongalactic astrophysics,’ ‘Earth and planetary astrophysics,’ ‘high energy astrophysical phenomena,’ ‘instrumentation and methods for astrophysics,’ ‘solar and stellar astrophysics’) and downloaded the .tex source file when possible from the Amazon S3 server (http://arxiv.org/help/bulk_data_s3) to determine the length of the paper and the number of equations and floats in the paper.
Figure 2 | Measured over predicted number of citations for papers authored by women. The error bars are obtained by bootstrapping and denote the 1σ errors on the mean of the measurement. The shaded green area shows the 1σ uncertainty on the best fit (see text for details). The dashed line indicates unity. The predictions of the citation numbers are based on non-gender-specific properties of the papers. The predictor is trained on the dataset containing papers authored by men only. We measure an average intrinsic bias of -10%, implying that women systematically receive -10% fewer citations than would be expected if they were men given the non-gender-specific properties of their papers.

Adding paper-specific information. We use the following procedure to determine the length and subfield for each paper. When the *.tex files are available, we run the tool TeXcount (http://app.uio.no/ifi/texcount/) with the default settings to obtain the number of words, floats, equations and mathematical expressions embedded in the text of each paper. The tool fails or measures only a very small number of words in the paper (<500) for some papers with multiple *.tex files associated with a single paper. We ignore the measurements for these papers in further analysis.

To estimate the subfield of the papers for which arXiv classification is not available, we train a random forest algorithm on the sample of papers for which both the field classification and an abstract are available. We are able to achieve a high accuracy of classification and find that about 80% of papers are correctly classified. Reassuringly, the misclassification is often between similar categories, such as between ‘cosmology and nongalactic astrophysics’ and ‘asteroseismology of stars’ or between ‘Earth and planetary atmospheres’ and ‘solar and stellar atmospheres’. If we exclude these similar misclassifications, then we find that the accuracy increases to ~90%. We then use this algorithm on all other papers to assign them to their field of research.

Adding author-specific information about institution. We determine the country of the institution of the first author to simplify and categorize the institutional information for each paper. In total, 85% of the papers include institutional information. We develop a list of about 100 keywords for which individual appearance in the affiliation string uniquely determines the country of origin. This list includes different spellings of country names, country codes, state names and abbreviations in the United States, and university and research institution names. Linking the affiliation strings to this list enables us to assign 97% of papers with affiliations uniquely to a country. To simplify this information, we assign the institutions to three categories: North America, Europe and other. We experiment with different classifications and find that these have a minimum effect on our conclusions.

Adding author-specific information about gender and seniority. Determining the first author's gender is complex because many authors publish using their initials instead of their full first names. We partially mitigate this problem by matching the first and last names with the initials of all authors from the dataset of all papers. In this way, we are able to determine the first name of an author even if they provide only initials in a particular paper, but use their first name at least once during their publishing career. We took special care to ensure bijection between the author information with initials only and the corresponding author information with a full first name. In many cases, the second and third first name (middle names) help to identify the unique full name provided by the initials. As a result of this methodology, we are able to uniquely identify different authors in the entire dataset and their reappearance.

We use the year of an author’s first-first-author paper as the baseline to define the seniority of an author. We define the seniority of an author as the number of years that have passed since their initial first-first-author publication. In cases where the exact first paper of the author cannot be identified due to possible confusion between authors with the same initials, we do not assign a seniority to such an author. In addition, we looked for authors who have changed their last name by looking for authors with last names that are part of other last names, while having the same first name. All possible cases have been individually checked to determine whether a change of last name is present. With this procedure, we are able to recover records for authors who have added another name to their surname during their publishing careers (perhaps due to marriage).

After determining the full first name, we match the name to three different databases to determine the gender. First, we look the name up with SexMachine (https://pyppi.python.org/pyppi/SexMachine/), a Python module. This database consists of 40,000 names from a wide geographical range that have been classified by native speakers. Second, we search for gender in the data available from the United States Social Security Administration and the UK Office of National Statistics, which track the gender of all children born in these countries (https://github.com/OpenGenderTracking/globalnamedata). It consists of about 100,000 names, but it does not have the geographical width of the first database. If the name is not found in these lists, we look the name up in Gender API (https://gender-api.com/), which includes nearly 2,000,000 names. If a given first name consists of several names, we check the gender for all of the names and weight the final gender assignment accordingly.

Cleaning and finalizing the dataset. The next step in our data processing is to remove those parts of the dataset that we judge to contain spurious information or for which we have incomplete information. In total, we downloaded 288,577 entries from ADS. We removed 58,836 entries (about 28%) from this initial dataset, giving a final dataset with 149,741 papers. The following ADS entries are removed: (1) entries with zero citations or zero references (4,417 ADS entries); (2) authors who have only published in Science and/or Nature (5,484 ADS entries); (3) entries with no specified author (491 ADS entries); (4) entries with no first name for the first author (such as collaboration articles: 7,713 ADS entries); (5) entries for which the first author used only initials for all publications available in the database (42,448 ADS entries); and (6) entries for which the gender of the first name of the first author could not be determined (2,260 ADS entries). Note that the number of the ADS entries removed for different individual reasons do not add up to the total number of removed entries as a result of overlaps. We verified that the gender distribution of the authors removed with cuts (1) and (2) is consistent with the gender distribution of the total sample. We removed entries by authors who have only published in Science and/or Nature because the initial database contains many papers with non-astronomical topics (such as geology and Earth composition). It is highly unlikely that authors from astronomy have only Science and/or Nature papers and have not been mentioned in any of the publications that specialized in astronomy. Supplementary Table 1 shows 12 randomly chosen lines from our dataset as an example.

Supplementary Fig. 1 shows the number of papers in our sample. The upper panel shows the number of published papers per year over time, the number of papers for which we were able to recognize the gender (in the main sample we discuss in this work) and the number of papers published by men and women. The lower panel shows the same information as a fraction of papers with recognized gender, papers authored by men or papers authored by women. We are able to recognize gender for a large fraction of papers, ranging from 60% in the 1960s and 1970s and increasing to 75–80% in the 1980s to 2010s. The fraction of recognized papers slightly decreased in the last few years as the fraction of authors who have published only a single paper or only a few papers has increased; for these authors, it is less likely that the full author name is available from one of their papers. We also note the slow, but constant, increase in the fraction of papers written by women, from <5% in the 1960s to ~25% in 2015. This trend is consistent with the overall increase in women faculty members in astronomy departments.

Completeness of seniority. We define seniority as the number of years since the author's initial first-author publication. As only papers after 1950 are included...
Received 9 September 2016; accepted 21 April 2017; published 26 May 2017; corrected 19 June 2017

References
1. Budden, A. E. et al. Double-blind review favours increased representation of female authors. Trends Ecol. Evol. 23, 4–6 (2008).
2. Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M. & Handelsman, J. Science faculty’s subtle gender biases favor male students. Proc. Natl Acad. Sci. USA 109, 16,474–16,479 (2012).
3. Wennerás, C. & Wold, A. Nepotism and sexism in peer-review. Nature 387, 341–343 (1997).
4. Conley, D. & Stadmuk, J. Gender matters: a call to commission more women writers. Nature 488, 590–590 (2012).
5. West, J. D., Jacquet, J., King, M. M., Correll, S. J. & Bergstrom, C. T. The role of gender in scholarly authorship. PloS ONE 8, e66212 (2013).
6. Ghiasi, G., Larivière, V. & Sugimoto, C. R. On the compliance of women engineers with a gendered scientific system. PloS ONE 10, e0145931 (2015).
7. Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. Bibliometrics: global gender disparities in science. Nature 504, 211–213 (2013).
8. Davenport, J. R. A. et al. Studying gender in conference talks — data from the 223rd Meeting of the American Astronomical Society. Preprint at http://arXiv.org/abs/1403.3091 (2014).
9. Pritchard, J. et al. Asking gender questions: results from a survey of gender and question asking among UK astronomers at NAM2014. Astron. Geophys. 58, 8–12 (2015).
10. Patat, F. Gender systematics in telescope time allocation at ESO. The Messenger 165, 2–9 (2016).
11. Reid, I. N. Gender-correlated systematics in HST proposal selection. Publ. Astron. Soc. Pacif. 126, 923–934 (2014).
12. Women, Minorities, and Persons with Disabilities in Science and Engineering (National Science Foundation, 2013).
13. Breiman, L. Random forests. Machine Learning 45, 5–32 (2001).
14. Hong, L. & Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proc. Natl Acad. Sci. USA 101, 16,385–16,389 (2004).
15. Medin, D. L. & Lee, C. D. Diversity makes better science. APS Observer (27 April 2012); www.psychologicalscience.org/observer/diversity-makes-better-science#.WQxB3psYo-aE
16. Page, S. E. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (Princeton Univ. Press, 2007).
17. Irie, R., White, S., Garrett, A. & Anderson, G. Women Among Physics and Astronomy Faculty (American Institute of Physics Statistical Research Center, 2013).
18. Pedregosa, F. et al. Machine learning in Python. J. Machine Learning Res. 12, 2825–2830 (2011).
19. Mathematica, Version 11.1 http://support.wolfram.com/kb/472 (Wolfram Research, 2017).
20. Breiman, L., Friedman J., Stone, C. J. & Olshen, R. A. Classification and Regression Trees (Chapman & Hall/CRC, 1984).
21. Hunter, J. D. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9, 90–95 (2007).

Acknowledgements
We thank J. Woo for giving detailed comments on the manuscript. We acknowledge the stimulating comments given to us by M. Urry, R. Schubert, R. Marino, B. Trakhtenbrot, I. Moose and E. Pourmarnas. We thank A. Bluck for proofreading the manuscript. We acknowledge support from the Swiss National Science Foundation. This research made use of the National Aeronautics and Space Administration’s Astrophysics Data System, the arXiv.org preprint server and the Python plotting library Matplotlib.

Author contributions
N.C. initiated the project and carried out the data analysis. S.T. created the name-matching algorithm and prepared the sample. S.B. created the algorithm that matched the authors with their geographical location. N.C. and S.T. wrote the paper. All authors discussed the results and commented on the manuscript.

Additional information
Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to N.C.

How to cite this article: Caplar, N., Tacchella, S. and Birrer, S. Quantitative evaluation of gender bias in astronomical publications from citation counts. Nat. Astron. 1, 0141 (2017).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests
The authors declare no competing financial interests.
Corrigendum: Quantitative evaluation of gender bias in astronomical publications from citation counts

Neven Caplar, Sandro Tacchella and Simon Birrer

Nature Astronomy 1, 0141 (2017); published 26 May 2017; corrected 19 June 2017.

In the version of this Letter originally published, the author’s name in ref. 11 was incorrect and should have read ‘Reid, I. N.’