nature ecology & evolution

Article

https://doi.org/10.1038/s41559-023-01999-w

Peer review perpetuates barriers for historically excluded groups

Received: 3 October 2022

Accepted: 31 January 2023

Published online: 13 March 2023

Check for updates

Olivia M. Smith[1,2,3], Kayla L. Davis[1,2], Riley B. Pizza[1,4], Robin Waterman[1,4], Kara C. Dobson[1,2], Brianna Foster[1,5], Julie C. Jarvey[1,2], Leonard N. Jones[1,2,6], Wendy Leuenberger[1,2], Nan Nourn[1,3,7], Emily E. Conway[1,4], Cynthia M. Fiser[1,5], Zoe A. Hansen[1,8,9], Ani Hristova[1,2], Caitlin Mack[1,2], Alyssa N. Saunders[1,2], Olivia J. Utley[1,2], Moriah L. Young[1,2] & Courtney L. Davis[1,2,10]

Peer review is central to the scientific process and scientists' career advancement, but bias at various stages of the review process disadvantages some authors. Here we use peer review data from 312,740 biological sciences manuscripts across 31 studies to (1) examine evidence for differential peer review outcomes based on author demographics, (2) evaluate the efficacy of solutions to reduce bias and (3) describe the current landscape of peer review policies for 541 ecology and evolution journals. We found notably worse review outcomes (for example, lower overall acceptance rates) for authors whose institutional affiliations were in Asia, for authors whose country's primary language is not English and in countries with relatively low Human Development Indices. We found few data evaluating efficacy of interventions outside of reducing gender bias through double-blind review or diversifying reviewer/editorial boards. Despite evidence for review outcome gaps based on author demographics, few journals currently implement policies intended to mitigate bias (for example, 15.9% of journals practised double-blind review and 2.03% had reviewer guidelines that mentioned social justice issues). The lack of demographic equity signals an urgent need to better understand and implement evidence-based bias mitigation strategies.

Peer review is a core part of the scientific process and vital for advancing scientists' careers. Yet peer review is not experienced similarly by all scientists[1], and may be negatively influenced by implicit or explicit biases based on author gender, geography, institution, race or other demographics[2–5]. Despite widespread concerns that peer review bias disadvantages scientists from historically excluded backgrounds[1,6], there have been no empirical studies that synthesize the extent of such biases across more than one demographic, nor solutions to mitigate bias. Accordingly, there remains debate on whether peer review bias is a substantial issue and, if so, how to combat it[7,8].

Demographic bias can manifest at any stage in the review process, from initial editorial decisions, to reviewer assessments or decisions

[1]Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, Michigan, USA. [2]Department of Integrative Biology, Michigan State University, East Lansing, MI, USA. [3]Center for Global Change and Earth Observations, Michigan State University, East Lansing, MI, USA. [4]Department of Plant Biology, Michigan State University, East Lansing, MI, USA. [5]Department of Entomology, Michigan State University, East Lansing, MI, USA. [6]Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA. [7]Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA. [8]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA. [9]Department of Plant Pathology, University of Minnesota, Saint Paul, MN, USA. [10]Cornell Laboratory of Ornithology, Cornell University, Ithaca, NY, USA. e-mail: olivia.m.smith@wsu.edu

https://doi.org/10.1038/s41559-023-01999-w

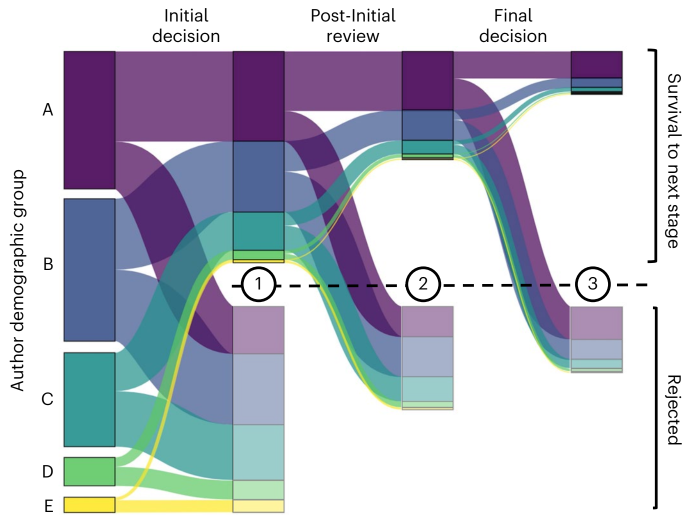

Fig. 1 | Bias can filter scientists from different demographics across the review process, leading to fewer published manuscripts by historically excluded groups. Some demographics submit more manuscripts than others. Initial editorial decisions can create a first filter by rejecting proportionally more manuscripts from some demographics due to bias (B–E; hypothetical historically excluded groups). If manuscripts are sent to review, reviewer bias can be reflected in worse reviewer recommendations, cascading into and compounding with editorial bias in decisions to reject manuscripts after the first round of review ('post-initial review'). This cumulative bias leads to a proportionally higher per cent of the dominant group's (A; for example, white males in high-income, English-speaking countries), and proportionally lower per cent of the non-dominant groups' (B–E) manuscripts being published. Numbers represent stages where the impacts of potential interventions to reduce bias would manifest (Supplementary Table 1). Examples of interventions that could theoretically reduce bias include triple-blind peer review, which could be beneficial at points 1 and 2; double-blind peer review, diversifying reviewer pools, providing clear reviewer guidelines and publishing reviews, which could be beneficial at point 2; and diversifying editorial boards, which could be beneficial at points 1, 2 and 3. Figure uses theoretical rather than actual data and is not based on specific demographic groups.

after review[4,5,9] (Fig. 1 and Extended Data Fig. 1). Yet, prior literature examining bias tends to focus on just one or two stages of the review process and considers few axes of demographic diversity (for example, ref. [10]). Thus, our understanding of where in the review process authors are filtered is limited (Fig. 1), especially considering the vast diversity of backgrounds the worlds' scientists represent.

There has long been heated debate on how to reduce disparate review outcomes, with no agreement on the best approach[11–13] (Supplementary Table 1). A notorious example is the lack of consensus on the efficacy of double-blind peer review at eliminating bias from the review process. Although some studies suggest that double-blind review can increase representation of authors from different backgrounds[14], others have argued against it because of inefficacy, hidden conflicts of interest and required changes to online submission systems (for example, refs. [8,14–19]). The efficacy of alternative peer-review models or other strategies aimed at reducing bias remains largely unknown, despite the pressing need to promote greater diversity amongst the scientific community and published works[20].

In this Article, we first conduct a meta-analysis to examine evidence for differential peer-review outcomes in the biological sciences at multiple stages of the review process based on demographics of the first, corresponding and last authors (for definitions, see Supplementary Table 2). We then evaluate the efficacy of proposed solutions to reduce bias, including double-blind review and editor/reviewer homophily. These analyses leverage data from 31 studies, which cumulatively examined 312,740 manuscripts submitted to >640 journals—including Nature Portfolio journals, *Science* and *Proceedings of the National Academy of Sciences*. Our dataset represents 4,529,971 author position/demographic/review stage interactions, which upon publication will be the largest publicly available dataset of this kind for future work to build upon. Finally, we describe the current landscape of peer review in the subfields of ecology and evolution by collecting peer review policy data from the websites of 541 journals. Altogether, we find that author demographics predict review outcomes; clear, evidence-based solutions to alleviate review bias are lacking; and relatively few journals are pro-actively combatting bias.

Results and discussion

Disparate peer review outcomes by author demographics

We found evidence for disparate peer review outcomes for all demographics that we examined at one or more stages in the review process (Supplementary Tables 3–40). We found the most data on outcomes by author assumed gender (Supplementary Data 1). Assumed female authors had worse or similar outcomes compared with assumed male authors, depending on the author position and review stage (Fig. 2).

We found notably lower success throughout the review process for authors with institutional affiliations in Asia, in countries where English is not a primary language and in countries with lower Human Development Indices (HDI; Figs. 2 and 3). Compared with authors with affiliations located in Europe, North America and Oceania, authors with affiliations located in Asia had the most consistent disparities, but authors in Latin America and Africa also often had worse review outcomes. When considering a country's continent, language and HDI in the same model, each was still important in predicting overall decisions for first and corresponding authors (but none was important for review scores; Extended Data Fig. 2). Further, we found that authors in countries where English is not a primary language were more likely to submit their manuscripts to a higher number of journals before acceptance than their peers. For the subset of data available at the individual country level, we also conducted analyses looking at the per cent of the population that speaks English, rather than the binary primary/not primary used by the literature we meta-analysed. In these analyses, English was still a highly important predictor of review outcomes (although in some cases HDI relationships changed; Supplementary Tables 36–40).

Despite the importance of geography in predicting review outcomes, we found few studies with geographical data (Fig. 2 and Supplementary Fig. 1). Future work should aim to fill this information gap because we saw large differences in review success based on continent, language and HDI for the review stages for which we had data (for example, overall decisions; Figs. 2 and 3). Likewise, we found few datasets examining review outcome gaps due to author prestige (individual and institutional; for how prestige was measured, see Supplementary Table 2), although prestige may be an important mediator in peer review outcomes based on the data we found (Fig. 2).

Possible solutions to reduce bias

We searched the literature for studies that evaluated potential solutions to mitigate bias but only found data on double- versus single-blind review models and editor/reviewer homophily (Supplementary Figs. 2 and 3). In some cases (for example, final decisions for assumed female authors), double-blind review reversed gender gaps in acceptance rates (Fig. 4a–c and Supplementary Tables 41–53). Double-blind review also appeared to 'level the field' for authors from countries with lower HDIs (Extended Data Fig. 3). That is, the difference in overall acceptance rates between authors in countries with lower versus higher HDIs was not as pronounced when authors opted for double-blind instead of single-blind review at the time of submission.

However, results on the impacts of double-blind peer review were far from consistent across review stages and demographics

https://doi.org/10.1038/s41559-023-01999-w

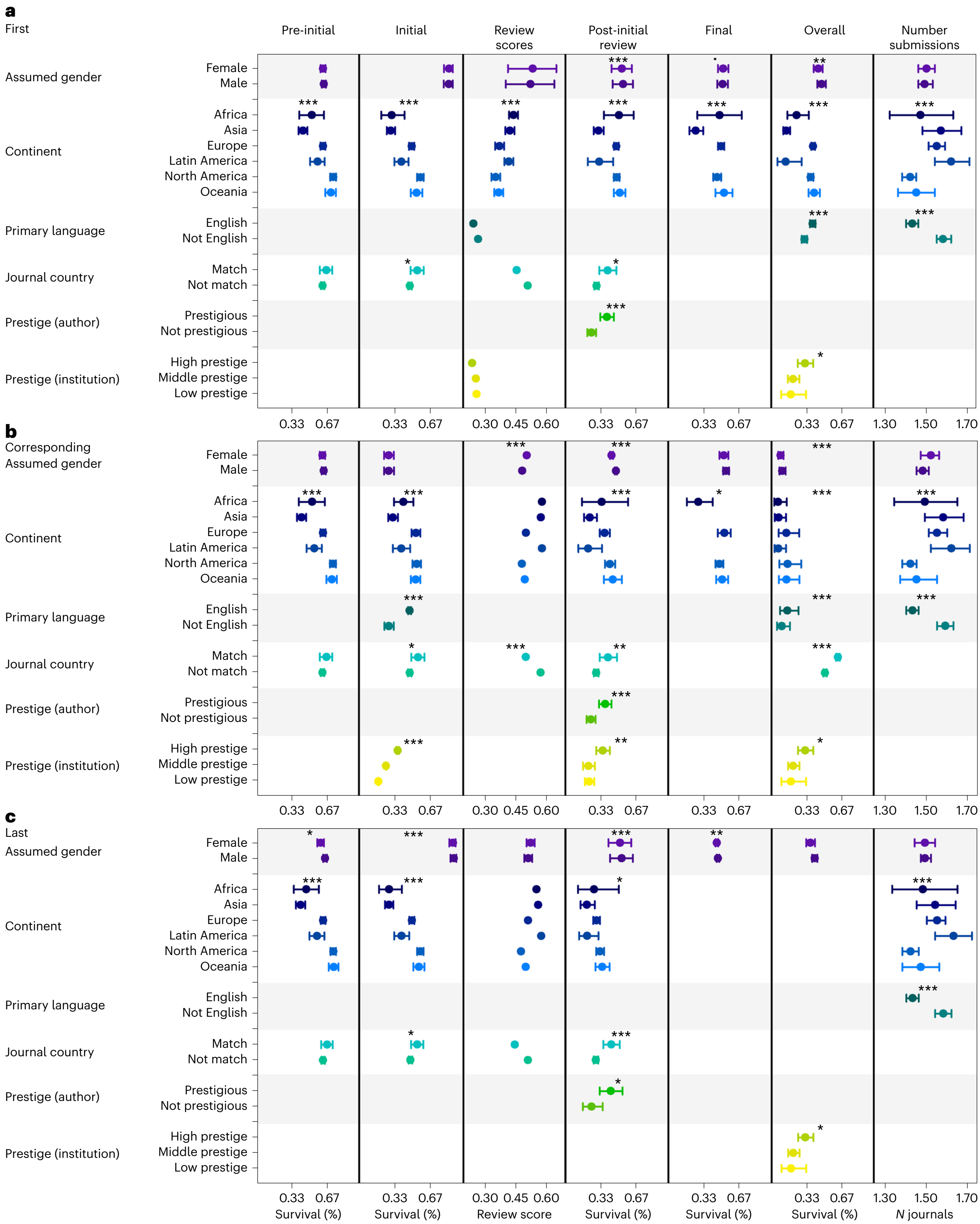

Fig. 2 | Author demographics predict review outcomes across the peer review process. a–c, Estimated means and 95% confidence intervals predicting review outcomes (pre-initial decisions, initial decisions, review scores (lower are better), post-initial review decisions, final decisions, overall decisions and number of journals submitted to before acceptance) by author demographics for first (**a**), corresponding (**b**) and last (**c**) author. Note that outcomes in each box will vary in scale because underlying studies typically focused on limited demographics/review stages in specific journals. Thus, the exact values in each box will vary, but inference can be made on outcome gaps between demographics across boxes. Review scores with data from one study per demographic category show the underlying mean. Colour used as a visual aid to track rows. $\cdot P < 0.10$, $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$ when reported for the estimated effects of demographic category on review outcomes (two-sided tests). For sample sizes, statistical tests used and exact P values, see Supplementary Tables 3–29.

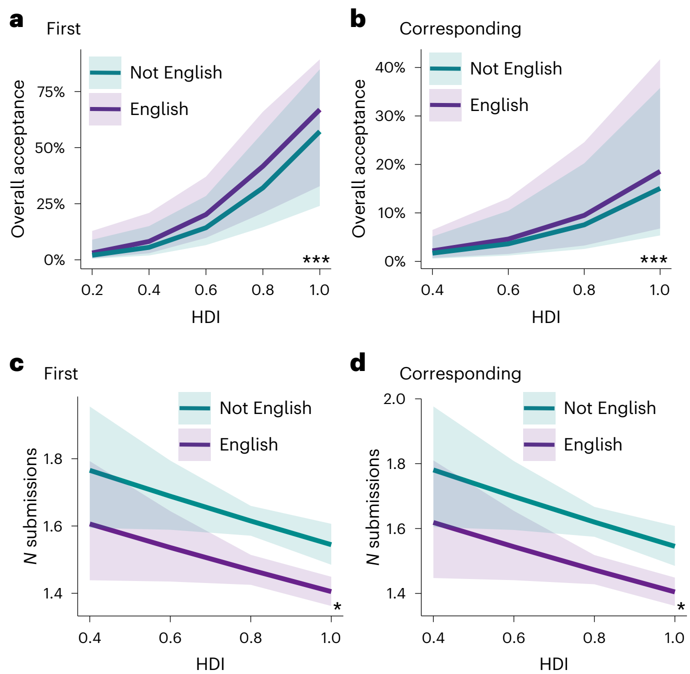

Fig. 3 | Authors in countries with lower HDI and where English is not a primary language have lower overall acceptance rates and submit their manuscripts to more journals before acceptance. a,**b**, Estimated means and 95% confidence intervals for first (**a**) and corresponding (**b**) author overall acceptance rate considering author continent, country primary language and HDI. **c**,**d**, Estimated means and 95% confidence intervals for number of journals submitted to before acceptance for first (**c**) and corresponding (**d**) author country primary language and HDI. ***$P < 0.001$ for both HDI and language (**a** and **b**); *$P < 0.050$ for HDI and $P < 0.001$ for language (**c** and **d**) (two-sided tests). For sample sizes, statistical tests used and exact P values, see Supplementary Tables 30 and 32.

(Fig. 4a–c and Extended Data Fig. 4). Further, although 225,249 manuscripts underlie our analyses on double-blind review, these data came from just four studies[10,21–23]. Only one of these studies experimentally allocated manuscripts to single- versus double-blind review[22], so we cannot disentangle differential outcomes based on self-selection of submitting single- versus double-blind manuscripts from true bias reduction. For example, we found that assumed female authors, when compared with assumed male authors, had a higher per cent of submissions sent for review when submitting double-blind manuscripts but a lower per cent sent for review when submitting single-blind manuscripts. The data underlying these analyses came from a study conducted on optional double-blind review in Nature Portfolio journals[21]. This suggests that there may be some self-selection bias when choosing a review model given that editors can see author information in double-blind review and make initial decisions. Future work experimentally allocating manuscripts to a review model, such as the one currently underway at *Functional Ecology*[24], can better determine the efficacy of double-blind review.

Another suggested solution to mitigate bias is to diversify editorial boards and reviewer pools. We found no data examining outcomes when author demographics 'match' editor demographics outside of assumed gender (Fig. 4d–f and Supplementary Tables 54–56). In this case, we generally found low evidence for editor gender homophily with two exceptions: higher final acceptance for female corresponding authors with female editors and a higher tendency towards the same for overall decisions.

We found more data examining potential reviewer homophily, but again, these data primarily focused on author assumed gender (Fig. 4g–i, Supplementary Fig. 4 and Supplementary Tables 57–70). We did not detect a strong homophily signal except for a few cases where reviewer and author demographics interacted (for example, assumed male first authors had higher acceptance at final decision with male reviewers and lower with female reviewers, when considering each author/reviewer interaction; Fig. 4g). The generally low evidence for homophily across the review process was surprising (for example, given experimental work[25]), and future studies should collect more data to assess the degree to which homophily may impact peer review. Given that we primarily found assumed gender homophily data, future work should examine possible homophily for other demographics (for example, geographical 'matching').

Current journal policy landscape

To describe the current landscape of peer review in the subfields of ecology and evolution, we collected peer review policy data from 541 journal websites (for a full overview, see Online Methods/Supplementary Data 2). Here we highlight patterns that emerged when grouping journals by impact factor (to compare society-affiliated versus unaffiliated journals, see Extended Data Figs. 5–7). No matter the breakdown, we see a similar story: journals across the board are taking few actions to reduce bias in peer review.

The five largest publishers in our dataset (Springer, Wiley, Elsevier, Taylor and Francis, and Oxford University Press) publish 57.3% of all ecology and evolution journals included (Fig. 5a and Extended Data Fig. 8). This suggests that consistent guidelines among these publishers could have widespread effects on peer review bias. While many journals suggest or require outside editing for authors for whom English is not a primary language, a mere 1.1% of journals offer free language editing (Fig. 5b). This might explain the lower acceptance rates for authors from countries where English is not a primary language (Figs. 2 and 3a,b), and rejection due to English grammar is commonly reported by journals and authors[26,27].

Across all impact factor categories, ~20% of journals recommend that authors suggest diverse reviewers (Fig. 5c), and those that do tend to focus on choosing reviewers from different geographic locations and institutions from the authors (Fig. 5d). Given the low evidence for homophily influencing peer review outcomes in our meta-analysis (Fig. 4 and Supplementary Fig. 4), this may not be a large source of bias in peer review. However, we found few data on homophily, and existing data were largely limited to gender and used assumed, rather than actual, gender. Questionnaire surveys that enable analysis of editor and reviewer demographics could better evaluate homophily effects.

Less than 20% of ecology and evolution journals were implementing alternative peer review models to single-blind review (Fig. 5e), although 22.7% of journals were not transparent about the review model on their website and were, therefore, assumed to be single-blind. High-impact journals were most likely to adopt double-blind and open review models (Fig. 5e). However, the percentage of high-impact journals using these practices was still small (30.3% did not use single-blind review models). High- and mid-impact journals were also most likely to publish reviews alongside articles, but it was still less than 13% for each (Fig. 5f).

Additionally, 56.7% of all journals did not have reviewer guidelines. Of the journals that did, many linked to publisher- rather than journal-specific policies (Fig. 5g) and did not mention issues related to social justice, such as implicit biases based on author assumed demographics or explicit comments regarding English language editing or errors (Fig. 5h). It is important to note that we were only able to gather journal policy data from publicly available websites. Some journals may, for example, provide additional reviewer guidelines via email or require authors to suggest reviewers during the submission process but not explicitly state this on their website. This lack of transparency is not only problematic for studies such as ours that aim to document

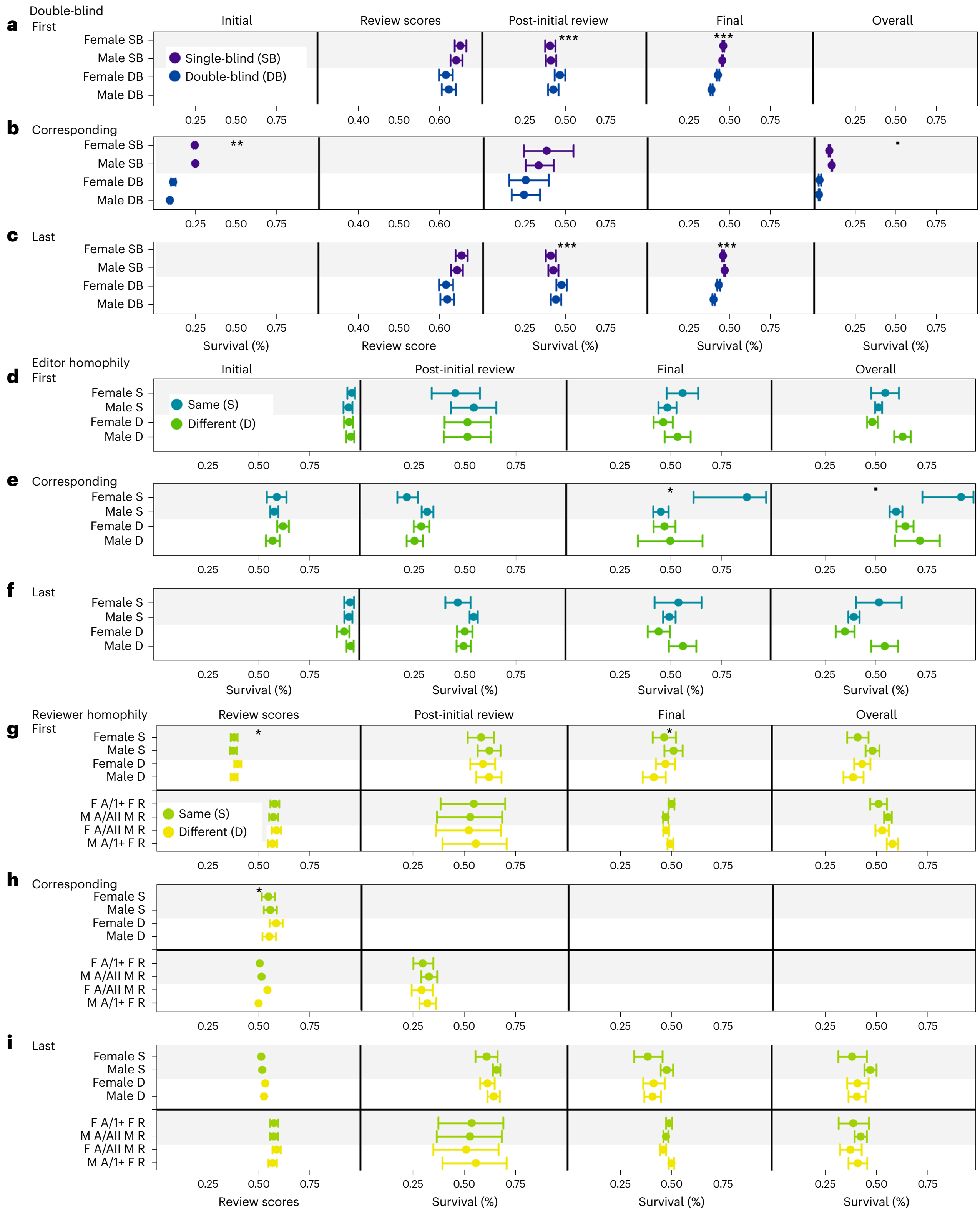

Fig. 4 | Double-blind review and editor/reviewer homophily can lead to differential outcomes by author assumed gender. a–i, Estimated means and 95% confidence intervals predicting review outcomes (initial decisions, review scores (lower are better), post-initial review decisions, final decisions and overall decisions) by author assumed gender, interacting with review model (**a**–**c**) (double- versus single-blind review) for first (**a**), corresponding (**b**) and last (**c**) authors; interacting with editor assumed gender (**d**–**f**) for first (**d**), corresponding (**e**) and last (**f**) authors; and interacting with reviewer assumed gender (**g**–**i**) for first (**g**), corresponding (**h**) and last authors (**i**). Studies examining gender homophily either classified reviewer gender using (1) each author/reviewer interaction (labelled female/male on axis) or (2) if a manuscript had all male versus 1+ female reviewers. F A/1 + F R = female author, 1+ female reviewer; M A/All M R = male author, all male reviewers; F A/All M R = female author, all male reviewers; M A/1 + F R = male author, 1+ female reviewer. Review scores with data from one study per review stage show the underlying mean. $\cdot P < 0.10$, $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$ when reported for the estimated effects of the interaction on review outcome (two-sided tests). For sample sizes, statistical tests used and exact P values, see Supplementary Tables 41–43 (**a**–**c**), 54–56 (**d**–**f**) and 57–62 (**g**–**i**) .

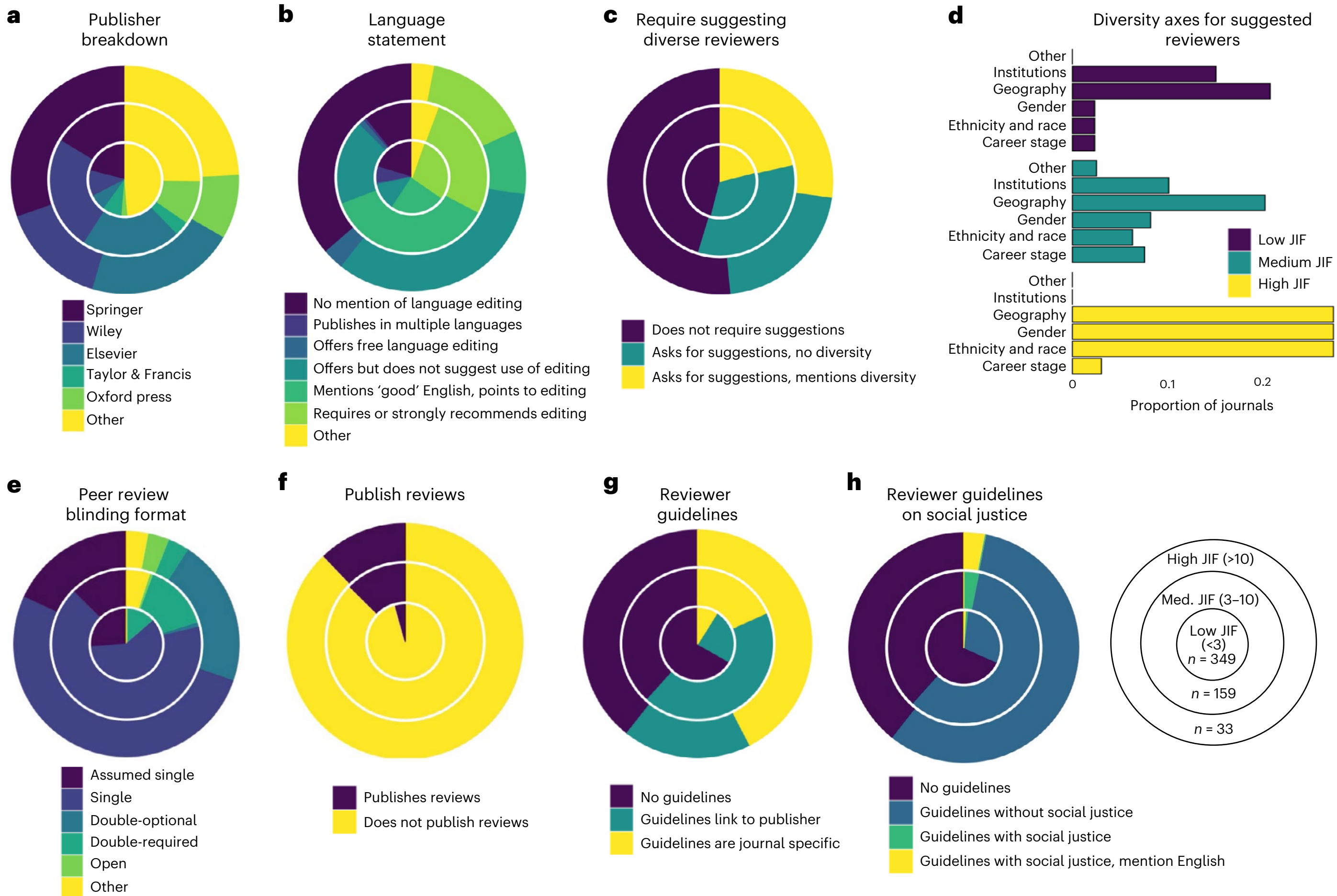

Fig. 5 | Ecology and evolution journals are taking a few actions to reduce bias in peer review. a, A few major publishers account for a large proportion of journals. **b**, Many journals require or suggest the use of paid language editing services for authors for whom English is not a primary language. **c**, Few journals prompt authors to suggest diverse reviewers on their websites. **d**, Of journals that do ask authors to suggest diverse reviewers on their websites, geography and institution were the most often mentioned axes of diversity (shows per cent of all journals that mentioned that axis of diversity, even if they did not prompt authors to suggest diverse reviewers). **e**, Single-blind review is the most common model. **f**, Only a small percentage of journals publish reviews alongside accepted manuscripts. **g**, Many journals do not have their own reviewer guidelines. **h**, Most reviewer guidelines do not mention social justice issues related to demographic equity, including the evaluation of English. Concentric circles show high journal impact factor (JIF) (outer-most ring, JIF >10, $n = 33$ journals), medium impact factor (JIF of 3–10, $n = 159$ journals) and low/no impact factor journals (inner-most ring, JIF <3, $n = 349$ journals).

current policies and bias mitigation strategies, but also for authors of various demographics that are looking for journals that will promote an objective review of their science.

Finally, we documented the geographic location of each journal as well as their Editor-in-Chief(s) (EICs) (Fig. 6, Extended Data Figs. 6, 7, 9 and 10 and Supplementary Fig. 5). We found that journals and editorial boards are concentrated in the geographic regions (Fig. 6a,b) with the highest overall acceptance rates (Fig. 2). This pattern is particularly evident in high-impact journals, where only a handful of countries are represented (Fig. 6c,d and Extended Data Fig. 10). These same countries (for example, the United States and the United Kingdom) also host a disproportionately high number of editorial board members and mid- to low-impact journals (Fig. 6e–h, Extended Data Figs. 9 and 10 and Supplementary Fig. 5). Although we were not able to directly examine editor homophily on the basis of country/region, we often found better outcomes for authors who submitted to journals from their own country (Fig. 2). Thus, the high concentration of journals and EICs in just a few regions could be problematic for authors from countries that are not as well represented, as suggested in our meta-analysis (for example, Fig. 2). Editorial board diversity beyond the EIC is also an important consideration. In our journal policy data collection, Elsevier journals were the only ones to provide editor gender data on some of their websites (Supplementary Data 3). Across the 22 of 60 Elsevier journals to provide this information, 28.8% of responding editors were women and 0% were non-binary or gender diverse. Editorial board geographic diversity was available on all Elsevier journal websites, and the mean number of countries represented in a single editorial board was 16.3 (range 3–36), while the mean editorial board size was 54.8 (range 3–218).

Study limitations and future directions

Several limitations should be considered in interpreting the results of our study. First, our meta-analysis examined observational data from real journal submissions that cannot control for any underlying differences in article quality or other attributes that facilitate acceptance at a given journal. Accordingly, we are not able to determine causality (that is, bias) over correlation (for example, article quality or fit for a journal). This means that, although our study documents often strongly differential peer review outcomes based on author demographics, we cannot parse the variance caused by bias per se from other factors (for example, operational peer review filtering papers that are substandard for any reason).

https://doi.org/10.1038/s41559-023-01999-w

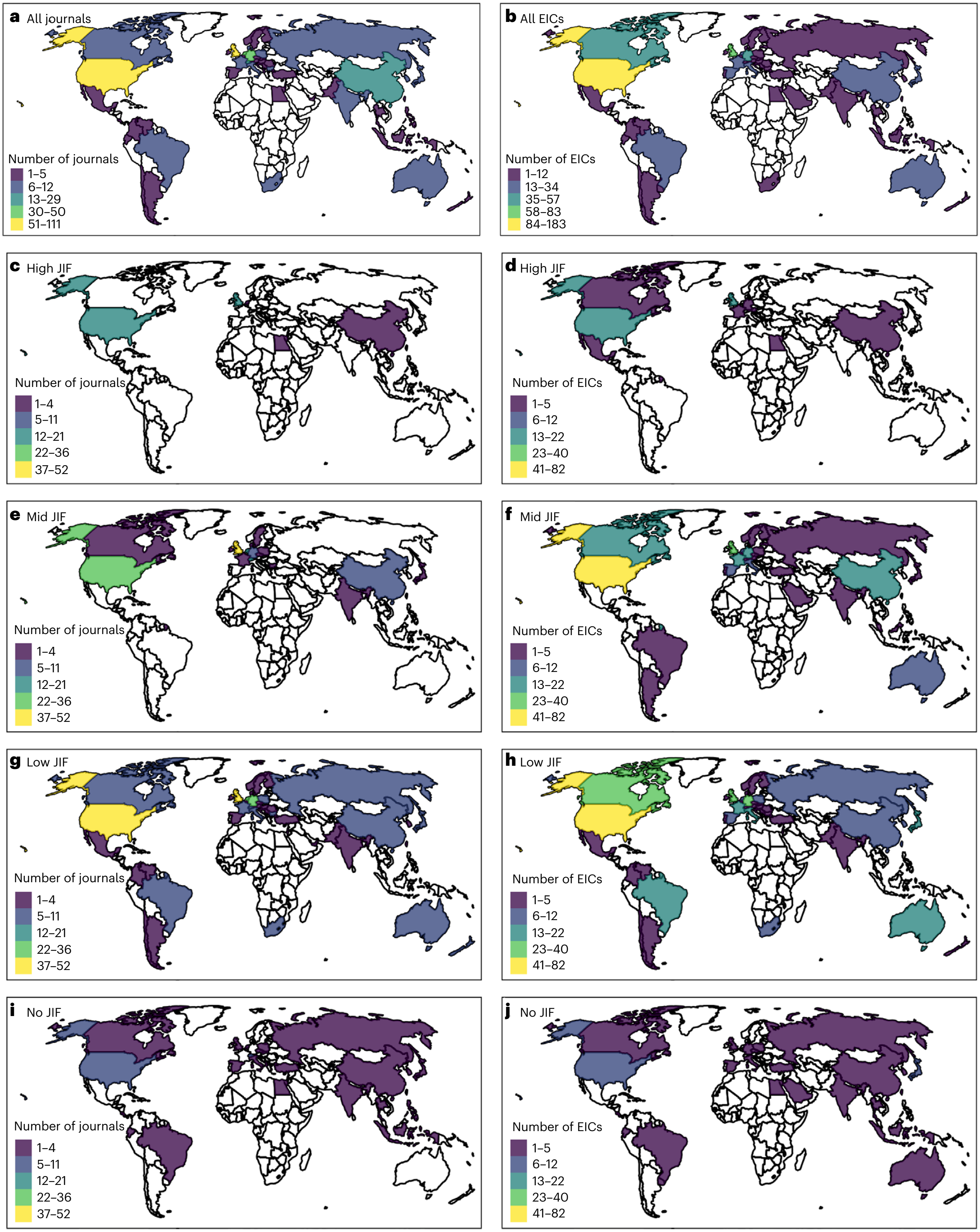

Fig. 6 | Journals and EICs are concentrated in few countries/regions. a,c,e,g,i, Journal locations for all journals (**a**), JIF >10 (**c**), JIF 3–10 (**e**), JIF <3 (**g**) and JIF not available (**i**). **b,d,f,h,j,** EICs' institutional affiliations' locations for all journals (**b**), JIF >10 (**d**), JIF 3–10 (**f**), JIF <3 (**h**) and JIF not available (**j**). JIF not available when no JCR have been indexed or has not been in JCR long enough. Warmer colours indicate more journals or EICs. Note that the total number of EICs exceeds the total number of journals because some journals have more than one EIC. For Europe inset, see Extended Data 10. Base maps were provided by the 'tmap' package in R (ref. [50]).

Experiments where the same manuscript is provided to editors and/or reviewers with different names or affiliations is one way to get around these limitations. However, we recovered only five studies that altered author details on the same manuscripts to see how assumed author demographics alter review outcomes (Supplementary Table 71). More experimental work should be done, but the studies we found suggested that demographic bias may impact review outcomes. For example, outcome disparities for authors with affiliations in countries with low HDI may be due to scientists there having fewer resources. Yet other experimental work has shown that people rate the same published abstracts lower when they are allegedly from a low-income country than when they are allegedly from a high-income country[28,29]. It is likely that our results partially reflect both disparity in access to scientific resources and bias. Journals can directly target the latter and should seek to understand what policies and practices best reduce outcome disparity, then implement them. Regardless of the variance caused by bias per se, our results should give the scientific community pause and momentum to rectify the systematic inequities we documented by rigorously evaluating the extent of disparate peer review outcomes and potential solutions to close gaps.

While a strength of our study is synthesizing prior work that individually tended to analyse just one or two stages of the review process and consider few demographics, our meta-analytic approach also causes a limitation: there were often few studies underlying any given analysis (Supplementary Figs. 1–3). Thus, conclusions on any given author position/demographic/review stage could be driven by studies with many datapoints. Additionally, our study is largely based on assumed author demographics (that is, gender using names and language using country primary language). This approach has the benefit that it may be more reflective of what editors and reviewers assume[30], but it is not reflective of outcomes by actual demographics. Further, there may be other negative impacts of peer review bias that we could not explore, such as changes in reviewer comment length/tone/professionalism, time of review, number of revisions and number of reviewers—we found few or no studies that provided data on these areas (Supplementary Figs. 6–9 and Supplementary Tables 3–70).

Finally, we note that, while we can meta-analyse data from prior studies, we do not have the lived experience of many of the issues documented. This work was inspired by witnessing explicitly biased comments as co-authors and co-reviewers. We also were inspired by our own seemingly worse peer review outcomes compared with our colleagues that do not identify with historically underrepresented groups. However, we emphasize that we do not have the authority to comment on identities or experiences outside our own.

Outlook

Our study reveals disparate outcomes for authors based on their demographics. Yet, solutions to mitigate bias were rarely studied, and actions taken by journals to mitigate bias were sparse. Future work needs to evaluate review outcomes across a wider range of demographics than has been done to date, and a wider range of solutions should be examined (for example, those proposed in refs. [8,31]; Supplementary Table 1). Data accessibility from journals is a major limitation to advancing our understanding of the extent of and solutions to bias[3,14]. We recommend that more journals collect peer review data to evaluate bias (for example, refs. [32]) and efficacy of solutions, then make these data available to scientists for rigorous evaluation and transparency. Only then can we have data-driven solutions to mitigate a serious problem. Scientists can also administer author questionnaire surveys (for example, refs. [1,5]) to circumvent data accessibility issues, which could provide a complementary and more comprehensive picture of peer review bias.

Broad solutions for journals and studies to consider include evaluating author outcomes across the full spectrum of review models from open, to single-, double- and triple-blind review[8]. More research is needed to determine which of these four general review models may be most effective given that we found no quantitative data examining how review model impacts outcomes by author demographics beyond double- versus single-blind review. In addition to redesigning review models, journals could implement and evaluate the efficacy of targeted diversity, equity and inclusion changes to try to make the process fairer. For an overview of ideas, see refs. [6,8,31] and Supplementary Table 1. Robust quantitative work is necessary before we can move the scientific community towards iron-clad bias mitigation strategies.

Peer review is a central part of the scientific process. Implicit bias in peer review restrains scientific advancement by rejecting or delaying publication of important works that are only perceived as lower quality. For example, we observed the worst review outcomes for authors from parts of the world that have notoriously few published studies, despite common calls for greater representation among published works. Reducing bias against scientists from these areas could advance the amount of published works from these underrepresented locations beyond just those of 'parachute researchers'[33]. Peer review is also critical to the advancement of scientists' careers, and any bias in peer review can stifle the productivity of underrepresented and historically excluded groups, maintaining the leaky pipeline[6]. We hope our synthesis will prompt future studies to swiftly and fully document these gaps and identify viable solutions for journals to implement. As scientists, we are well suited to the task of collecting and analysing data to solve complex problems. The time to use these talents to build a less biased peer review system is now.

Methods

Meta-analysis of disparate outcomes in peer review

Literature search. We searched for studies examining the role of author demographics in journal article peer review evaluations and outcomes using the Web of Science (last search January 2022). To establish our final systematic search terms, we conducted a series of preliminary searches to limit irrelevant results while finding commonly used words that we had initially missed. Our systematic Web of Science search required the following search terms: [TITLE = ('peer review*' OR 'review' OR 'reviewer*' OR 'referee*')] AND [TITLE = ('bias*' OR 'gender*' OR 'female' OR 'race*' OR 'racism' OR 'ethnicit*' OR 'English*' OR 'language' OR 'disabilit*' OR 'institution*' OR 'geography' OR 'socioeconomic*' OR 'blind' OR 'underrepresented')] AND [TOPIC = ('journal*' OR 'publication*')]. Our search yielded 2,202 studies whose titles and abstracts were then screened for suitability.

We determined suitability using seven inclusion criteria: (1) the study evaluated the impact of author demographics on article evaluations and outcomes during peer review; (2) the study focused on journal article review and not review of grant proposals, conference abstracts or posters, and so on; (3) the article was a primary study. However, non-primary studies were collected to screen for additional references; (4) the study concerned peer review of articles in the biological sciences. We defined the biological sciences following section/topic classifications in *Nature* and *Proceedings of the National Academy of Sciences* (for example, agricultural sciences, biochemistry, bioinformatics, cell biology, ecology/evolution, genetics, medical sciences, physiology, plant biology, psychology and systems biology); (5) data were not duplicated in another included study and could be extracted from the article or supporting data, or were sent by the authors. For studies examining acceptance/rejection rates, we further required that we could determine the total number of papers accepted/rejected by demographic category to model outcomes using a binomial distribution; (6) the study examined actual peer review scenarios of real articles linked to specific journals (anonymized or identified). However, we also collected and summarized studies that conducted experiments using fabricated manuscripts using realistic peer review scenarios; and (7) the study was published in a peer-reviewed journal (for example, not on a preprint server).

Of the 2,202 articles returned by our Web of Science search, 80 articles appeared to meet these inclusion criteria on the basis of their titles/abstracts and were read in full to determine if they did indeed meet all criteria. To minimize the chances of our Web of Science search missing suitable literature, we sought additional studies by screening the references in the 80 articles from Web of Science, resulting in an additional 142 articles that were then read in full to determine their suitability for inclusion. We did a final search through references in the 142 articles gathered from the exhaustive search, resulting in an additional 55 articles that were read in full to determine if they met our criteria. Finally, when emailing authors for data that could not be collected from their articles, we were sent three additional studies that were screened for inclusion. In total, we screened 280 articles by reading their full texts, of which 32 met all inclusion criteria to be included in our quantitative meta-analysis. However, after collecting data on all authorship positions provided by each article, we further required that studies provided data on first, corresponding and/or last authors. This was due to the large focus on these positions and to reduce the number of single effect sizes. One article (Stossel and Stossel[34]) was removed due to this additional criterion, leaving 31 used in analyses (Supplementary Data 4).

Five additional experiments that failed to meet our actual peer review scenarios criterion (6) have been summarized in Supplementary Table 71. Experimental studies that manipulate author details on the same manuscript can get at causality (that is, bias) and so provide valuable information. However, we did not find enough of these studies to meta-analyse them, especially given that they all had different foci, so we ultimately decided to summarize them. Additionally, we found four articles (Supplementary Table 72) that aimed to test the efficacy of double-blind peer review at reducing bias by examining the number of published papers by demographic group in paired single-blind/double-blind peer review journals and/or the number of published papers before/after double-blind peer review implementation. These papers did not examine outcomes during peer review, and thus failed to meet our inclusion criterion (1). However, given that one of these studies (Budden et al.[14]) has been highly cited and influential in arguing both for and against double-blind peer review, we summarized results from all four in Supplementary Table 72.

Variables collected for meta-analysis. One author (O.M.S.) collected data from studies included in our meta-analysis that assessed peer review outcomes by authors of different demographics in absence of intervention (that is, examining review disparity problems) and/or data that assessed peer review outcomes with an intervention (that is, examining possible solutions to mitigate review disparities). Interventions examined by studies in our meta-analysis included double-blind versus single-blind peer review and outcomes for authors based on homophily (that is, if their demographics 'matched' the editors and/or reviewers). We separated data by peer review stage (Extended Data Fig. 1 and Supplementary Table 2) including pre-initial decision, initial decision, review scores, post-initial review decisions, final decisions and overall decisions. We also collected data on time associated with different review stages, number of reviewers and number of revisions. Finally, one study looked at the number of submissions a manuscript went through before acceptance[5].

We separated peer review outcome data by review stage, author demographics (for example, assumed gender, continent and institutional prestige) and authorship position (that is, first, corresponding and last). The 31 included studies cumulatively examined 312,740 manuscripts submitted to >640 journals and represented 4,529,971 author position/demographic/review stage interactions. Studies that examined more than one demographic or authorship position typically had missing data for some author position/demographic/review stage combinations, so we used the largest single sample size of all possible author position/demographic/review stage interactions from each study to calculate the total manuscripts above.

For studies that provided data by an author's institution's country, we assigned continent and English as a primary language using the Central Intelligence Agency (CIA) World Factbook[35] following Burns and Fox[4]. However, the CIA World Factbook uses South America and North America, but we split the Americas into Latin America and North America[4]. We assigned English as a primary language of a country if English was the first listed language or an official language[4]. We used this binary classification because all studies that provided peer review data for language did so. However, for data on review outcomes by an author's institution's country, we were also able to assign the per cent of the population that speaks English following the procedures of Amano and Sutherland[36]. That is, we estimated the per cent of the population that speaks English from the CIA World Factbook[35], Ethnologue[37], *The Cambridge Encyclopedia of Language*[38] and the Eurobarometer Survey[39]. Ethnologue gives the total number of English speakers, so we divided the Ethnologue estimates by the population sizes in the CIA World Factbook. We used the maximum percentage from the four sources in our analyses (Supplementary Data 5). Finally, each country was assigned a HDI value using the closest available year to the mean study year (the HDI database starts in 1990).

For each study included in the meta-analysis, we (O.M.S., R.B.P. and R.W.) collected study attribute data including mean year data were collected by authors (Supplementary Fig. 10), journal(s) included within studies and if studies focused on journals in the health sciences and/or life sciences. Our study included at least 640 unique journals and probably more because one study[10] provided suitable data from 79 anonymous journals from which we could not determine overlap with other studies and another provided data from one anonymous journal[40]. In some cases, studies examined journals inside and outside of the biological sciences. We only used data from journals in the biological sciences where possible. If separating the data was not possible, we required that >70% of data were from journals in the biological sciences. For each journal, we collected journal impact factors (JIF) from Clarivate Journal Citation Reports (JCR) for the mean year data were collected. In some cases, the studies were conducted before 1997 (when Clarivate electronic entries begin), so we used the closest year for which we could find printed reports (1991 and 1994). In other cases, journals were not JCR indexed until after the mean study year, so we used the closest year that was indexed. In other cases (Squazzoni et al.[10]), the study did not provide the exact journals studied but did provide impact factors, so we used the study-provided values. Finally, in one case (Walker et al.[41]), the study examined peer review in Frontiers journals but did not provide exact journals. We searched Clarivate for all Frontiers journals indexed for the mean study years (2009–2010) and used the mean JIF for those.

Ecology and evolution journal policy assessment

Ecology and evolution journal selection. Candidate ecology and evolution journals were collected using a combination of (1) general web searches, (2) category searches across large publishers (Wiley, Elsevier, Springer, Sage Publications, Taylor and Francis, Oxford University Press, Cambridge University Press and University of Chicago Press) and (3) Clarivate's JCR. Journals under the following JCR categories were retrieved: Agriculture, Dairy and Animal Sciences; Agronomy; Biochemistry and Molecular Biology; Biology and Biogeochemistry; Environmental Sciences; Evolutionary Biology; Genetics and Heredity; Marine and Freshwater Biology; Microbiology; Plant Sciences; and Zoology. We found 1,877 journals that met these initial criteria.

To select our final list of journals for which we collected policy data, we first eliminated journals that were irrelevant to ecology and/or evolution or only very rarely publish papers on ecology and/or evolution. Excluded journal foci included animal, crop or range management journals (largely categorized under agronomy, agriculture, animal

 https://doi.org/10.1038/s41559-023-01999-w

science and environmental science/management); microbiology and molecular/cell biology journals with a focus on developmental mechanisms/processes or medical focus/clinical application; and 'system specific' journals with narrower organismal focus than botany and zoology (for example, mammalogy and herpetology).

We then screened the remaining 677 journals for inclusion by checking their journal Aims and Scope (or equivalent section under a different name). Two people independently scored each journal for (1) explicit mention of ecology and/or evolution (including mention of subdisciplines such as 'disease ecology' or 'trophic ecology') and (2) a focus on a single taxon or specific organismal system. If we were unable to determine suitability based solely on the Aims and Scope section, we looked at article foci in recent issues.

Any disagreements between the two scores were independently rescored by a third person, who also assigned journals to Tier 1 through Tier 4. Tier 1 journals were those that were explicitly ecology and/or evolution focused with a broad taxonomic scope (for example, *Ecology*, *Evolution* and *Journal of Applied Ecology*). Tier 2 journals were those that had a broad focus but included many ecology/evolution papers or had ecology/evolution subsections (for example, *Nature*, *Science* and *Proceedings of the National Academy of Sciences*). Tier 3 journals were marginally focused on ecology/evolution and/or were narrow in taxonomic/organismal system scope (for example, *Arctic* and *Environmental Entomology*). Tier 4 journals were clearly not ecology and/or evolution focused and were eliminated from all subsequent analyses. In total, there were 555 journals that were assigned to Tiers 1–3 for which we tried to collect journal policy data. We excluded several journals because they required an account to access author guidelines, had invalid websites or links to author guidelines, were no longer receiving submissions or were a book series. Our final dataset included 541 journals (Supplementary Data 2). Our search aimed to gather a comprehensive and representative list of general ecology and evolution journals rather than build a complete list of journals that occasionally publish ecology and evolution papers.

Ecology and evolution journal policy data collection. We used Clarivate's JCR to gather each journal's country, publisher, publishing language(s) and 2020 JIF. If publishers were subsidiaries of larger publishers, we rescored them to the largest publisher for our main-text summaries. If journals were not JCR indexed ($n = 23$), we collected information from the journals' websites. To collect journal policy data, one person initially scored each journal by checking journal websites. Then, two authors (O.M.S. and C.L.D.) rescored the data for consistency, sorted by publisher.

We collected data from the website(s) of each journal pertaining to (1) the journal's article language requirements; (2) language editing services available to authors; (3) journal prompts for authors to suggest diverse reviewers and what types of diversity authors were prompted to focus on in suggestions; (4) if the journal had reviewer guidelines at all, and if so, if they were specific to the journal or linked to a general publisher website (to be counted, reviewer/referee guidelines must have been listed as such, either in a separate webpage/document or as part of the author guidelines); (5) if reviewer guidelines mentioned social justice issues, and in particular, behaviours towards authors for whom English is a non-primary language; (6) what country the EICs' primary institution was in; (7) if the journal was affiliated with a society; (8) what peer review model (single-, double- or triple-blind; open; or other) the journal used; and (9) if the journal publishes referee reports with manuscripts. Elsevier journals also provide the total number of editors, number of countries represented in the editorial board and gender diversity from survey data, which we collected when available.

We classified a journal's peer review model as single-blind when the author identities were revealed to reviewers, but the reviewer identities remained unknown to the authors. If journals specified that reviewers could sign reviews, we still counted the review model as single-blind. Some journals did not specify a review model and were scored as 'assumed single-blind' if the title page with author details was part of the main manuscript document. We classified a peer review model as double-blind when both the reviewer and author identities remained unknown to each other. We further separated double-blind review into (1) optional double-blind wherein the journal allows the authors to choose to remain anonymous or be known to reviewers and (2) required double-blind wherein the journal requires authors to remain anonymous to reviewers. We defined triple-blind review as the author identities being unknown to the editors until after the initial decision and the author and reviewer identities being unknown to each other throughout the review process. Finally, we defined open review as the mandatory identification of reviewers at any point during the review or publication process. There were 11 journals in which the review models did not fit well within these definitions (for example, have public commentary period on a preprint server) that were marked as 'other'. In several cases, we emailed journals' editorial offices because the websites had contradictory statements on review model or no information indicating review model (for example, no manuscript formatting guidelines).

Statistical methods

We examined if demographic categorizations predicted review outcomes for each unique author position/review stage/demographic category we had data for using the package 'glmmTMB' in Program R v4.0.3 (refs. [42,43]). We used a binomial distribution for review stages with binary outcomes (that is, manuscripts survived that stage or were rejected, which included pre-initial, initial, post-initial review, final and overall decisions), a beta distribution for review score data, a Gaussian distribution for models examining number reviewers and number revisions, and a Poisson distribution for models examining number of journals submitted to before acceptance. We checked model assumptions using the 'DHARMa' package in R (ref. [44]).

When data came from more than one study for a given comparison, we accounted for the impact factors of journals used in the study, mean data collection year and other differences by (1) including study as a fixed effect when there were only two studies or (2) including JIF and year as fixed effects and study as a random effect when there were three or more studies. We checked for multicollinearity issues for models that included JIF and year using the variance inflation factor in the 'performance' package in R (ref. [45]). In the few cases where variance inflation factor was high (>5), we compared models that included either JIF or year using Akaike information criterion$_c$ in the 'bbmle' package in R (ref. [46]). We then used the model with the lower Akaike information criterion$_c$ value to make inference.

For demographic groups with data in three or more categories (that is, institutional prestige and continent), we first assessed if variables improved model fit using likelihood ratio tests. If $P < 0.05$, we used generalized Tukey's honest significant difference tests in the 'multcomp' package in R (ref. [47]) to determine which categorical variables differed. We used the 'emmeans' package in R (ref. [48]) to get estimated means and 95% confidence intervals, which were used in Figs. 2 and 4, Extended Data Fig. 4 and Supplementary Figs. 4 and 6–9 (Supplementary Data 6).

For review stages for which we had data at the individual country level (initial decisions for corresponding, review scores for first/corresponding/last, overall decisions for first/corresponding and number submissions for first/corresponding/last), we ran additional models that included each country's continent, language and HDI. We conducted one set of analyses using a binary fixed effect of English as a primary/not primary language for that country, which is how all included studies analysed outcomes by authors' assumed primary language. We then repeated all analyses using the per cent of the population that speaks English. We included JIF, year and study using the same protocol described above. For data at the individual country level, we checked for multicollinearity between continent, language and HDI as described above. Where multicollinearity was not an issue, we ran models including the additive effects of continent, language and HDI. Where multicollinearity was an issue, we ran models examining the impact of each of the three variables alone if all were too correlated to be in one model (that is, initial decisions for corresponding authors) or with all possible pairs if only all three were too correlated to be in the same model (that is, number of submissions for all positions and review scores for all positions when using per cent of the population that speaks English).

We repeated our procedures described above for models examining the ability of interventions (double-blind versus single-blind review; editor and reviewer homophily) to alter peer review gaps by testing for an interaction between author demographics and the intervention. To do so, we conducted likelihood ratio tests. For models using data at the individual country level (available for double-blind/corresponding author), we took a stepwise approach, testing the interaction between blinding format and each variable (continent, language and HDI; analysis set run once for language as a binary variable and once for per cent of the population that speaks English) sequentially in the otherwise fully additive model.

Outcomes in our meta-analysis are on different scales between each author position/demographic/review stage combination. This is because underlying studies were typically conducted on specific demographic groups for specific review stages at specific journals, and these underlying journals differ in their rejection rates and rigour. Thus, the exact values at any given author position/demographic/review stage combination will vary depending on the underlying studies. Accordingly, we can make inference about disparate outcomes at each author position/demographic/review stage, but the exact percentages/scales are not meaningful between author positions/demographics/review stages.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets generated and analysed during the current study are available on GitHub at https://github.com/CourtneyLDavis/Peer-Review-Perpetuates-Barriers. Data are also archived at Figshare (https://doi.org/10.6084/m9.figshare.21865830)[49]. Additional datasets that support the findings of this study are available from the United Nations Development Programme (the HDI; https://hdr.undp.org/en/content/download-data), The CIA World Factbook (continent, language; https://www.cia.gov/the-world-factbook/), Ethnologue (language; https://www.ethnologue.com/), The European Commission's Eurobarometer Survey (language; https://data.europa.eu/data/datasets/s1049_77_1_ebs386?locale=en), *Cambridge Encyclopedia of Language*[38] and Clarivate Journal Citation Reports (journal attributes; https://jcr.clarivate.com/jcr/home).

Code availability

R scripts used to analyse data and generate figures during the current study are available on GitHub at https://github.com/CourtneyLDavis/Peer-Review-Perpetuates-Barriers.

References

1. Silbiger, N. J. & Stubler, A. D. Unprofessional peer reviews disproportionately harm underrepresented groups in STEM. *PeerJ* **7**, e8247 (2019).
2. Peters, D. P. & Ceci, S. J. Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain Sci.* **5**, 187–195 (1982).
3. Tregenza, T. Gender bias in the refereeing process? *Trends Ecol. Evol.* **17**, 349–350 (2002).
4. Burns, C. S. & Fox, C. W. Language and socioeconomics predict geographic variation in peer review outcomes at an ecology journal. *Scientometrics* **113**, 1113–1127 (2017).
5. Fox, C. W. & Paine, C. E. T. Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecol. Evol.* **9**, 3599–3619 (2019).
6. Llorens, A. et al. Gender bias in academia: a lifetime problem that needs solutions. *Neuron* **109**, 2047–2074 (2021).
7. Ceci, S. J. & Williams, W. M. Understanding current causes of women's underrepresentation in science. *Proc. Natl Acad. Sci. USA* **108**, 3157–3162 (2011).
8. Brodie, S. et al. Equity in science: advocating for a triple-blind review system. *Trends Ecol. Evol.* **36**, 957–959 (2021).
9. Helmer, M., Schottdorf, M., Neef, A. & Battaglia, D. Gender bias in scholarly peer review. *eLife* **6**, e21718 (2017).
10. Squazzoni, F. et al. Peer review and gender bias: a study on 145 scholarly journals. *Sci. Adv.* **7**, eabd0299 (2021).
11. Godlee, F. & Dickerson, K. in *Peer Review in Health Sciences* (eds Jefferson, T., & Godlee, F.) 91–117 (Wiley, 2003).
12. Haffar, S., Bazerbachi, F. & Murad, M. H. Peer review bias: a critical review. *Mayo Clin. Proc.* **94**, 670–676 (2019).
13. Shoham, N. & Pitman, A. Open versus blind peer review: is anonymity better than transparency? *BJPsych Adv.* **27**, 247–254 (2021).
14. Budden, A. E. et al. Double-blind review favours increased representation of female authors. *Trends Ecol. Evol.* **23**, 4–6 (2008).
15. Whittaker, R. J. Journal review and gender equality: a critical comment on Budden et al. *Trends Ecol. Evol.* **23**, 478–479 (2008).
16. Budden, A. E. et al. Response to Whittaker: challenges in testing for gender bias. *Trends Ecol. Evol.* **23**, 480–481 (2008).
17. Webb, T. J., O'Hara, B. & Freckleton, R. P. Does double-blind review benefit female authors? *Trends Ecol. Evol.* **23**, 351–353 (2008).
18. O'Hara, B. Double-blind review: let diversity reign. *Nature* **452**, 28 (2008).
19. Naqvi, K. R. Double-blind review: the paw print is a giveaway. *Nature* **452**, 28 (2008).
20. Maas, B. et al. Women and Global South strikingly underrepresented among top-publishing ecologists. *Conserv. Lett.* **14**, e12797 (2021).
21. McGillivray, B. & De Ranieri, E. Uptake and outcome of manuscripts in Nature journals by review model and author characteristics. *Res. Integr. Peer Rev.* **3**, 1–12 (2018).
22. Alam, M. et al. Blinded vs. unblinded peer review of manuscripts submitted to a dermatology journal: a randomized multi-rater study. *Br. J. Dermatol.* **165**, 563–567 (2011).
23. Jagsi, R. et al. Attitudes toward blinding of peer review and perceptions of efficacy within a small biomedical specialty. *Int. J. Radiat. Oncol. Biol. Phys.* **89**, 940–946 (2014).
24. Fox, C. W. et al. Double-blind peer review—an experiment. *Funct. Ecol.* **33**, 4–6 (2019).
25. Lloyd, M. E. Gender factors in reviewer recommendations for manuscript preparation. *J. Appl. Behav. Anal.* **23**, 539–543 (1990).
26. Ramírez-Castañeda, V. Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: the case of Colombian researchers in biological sciences. *PLoS ONE* **15**, 1–15 (2020).
27. Ehara, S. & Takahashi, K. Reasons for rejection of manuscripts submitted to *AJR* by international authors. *Am. J. Roentgenol.* **188**, 113–116 (2007).
28. Harris, M., Macinko, J., Jimenez, G., Mahfoud, M. & Anderson, C. Does a research article's country of origin affect perception

of its quality and relevance? A national trial of US public health researchers. *BMJ Open* **5**, 1–10 (2015).
29. Harris, M. et al. Explicit bias toward high-income-country research: a randomized, blinded, crossover experiment of English clinicians. *Health Aff.* **36**, 1997–2004 (2017).
30. Steinpreis, R. E., Anders, K. A. & Ritzke, D. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: a national empirical study. *Sex. Roles* **41**, 509–528 (1999).
31. Mehta, D. et al. Ways to increase equity, diversity and inclusion. *eLife* **9**, 1–4 (2020).
32. Else, H. & Perkel, J. M. The giant plan to track diversity in research journals. *Nature* **602**, 566–570 (2022).
33. Stefanoudis, P. V. et al. Turning the tide of parachute science. *Curr. Biol.* **31**, R184–R185 (2021).
34. Stossel, T. P. & Stossel, S. C. Declining American representation in leading clinical-research journals. *N. Engl. J. Med.* **322**, 739–742 (1990).
35. The World Factbook. *Central Intelligence Agency* https://www.cia.gov/ (2021).
36. Amano, T. & Sutherland, W. J. Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. *Proc. R. Soc. B* **280**, 20122649 (2013).
37. Eberhard, D. M., Simons, G. F. & Fennig, C. D. *Ethnologue: Languages of the World* (SIL International, 2022).
38. Crystal, D. *The Cambridge Encyclopedia of Language* (Cambridge Univ. Press, 2010).
39. Special Eurobarometer 386: Europeans and their languages. *European Commission* https://data.europa.eu/data/datasets/s1049_77_1_ebs386?locale=en (2014).
40. Zhang, X. Effect of reviewer's origin on peer review: China vs. non-China. *Learn. Publ.* **25**, 265–270 (2012).
41. Walker, R., Barros, B., Conejo, R., Neumann, K. & Telefont, M. Personal attributes of authors and reviewers, social bias and the outcomes of peer review: a case study. *F1000Research* **4**, 21 (2015).
42. Brooks, M. E. et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R. J.* **9**, 378–400 (2017).
43. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2020).
44. Hartig, F. DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.4.4 http://florianhartig.github.io/DHARMa/ (2021).
45. Ludecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P. & Makowski, D. performance: an R package for assessment, comparison and testing of statistical models. *J. Open Source Softw.* **6**, 3139 (2021).
46. Bolker, B., & R Core Team. bbmle: tools for general maximum likelihood estimation. R package version 1.0.24 https://CRAN.R-project.org/package=bbmle (2021).
47. Hothorn, T., Bretz, F. & Westfall, P. Simultaneous inference in general parametric models. *Biom. J.* **50**, 346–363 (2008).
48. Lenth, R. V. emmeans: Estimated marginal means, aka least-squares means. R package version 1.7.1-1 https://CRAN.R-project.org/package=emmeans (2021).
49. Smith, O. M. et al. Peer review perpetuates barriers for historically excluded groups. *figshare* https://doi.org/10.6084/m9.figshare.21865830 (2023).
50. Tennekes, M. tmap: thematic maps in R. *J. Stat. Softw.* **84**, 1–39 (2018).

Acknowledgements

We thank authors who sent us additional data we requested (fully documented in Supplementary Data 4), especially S. Burns, L. Fabrigar, M. Flemming, C. Fox, B. McGillivray, T. Paine, R. Petty and T. Tregenza. V. Pan assisted with data collection. A. Hayes and Z. Katz provided input on data collection. Y. Ge, L. Rasmussen and J. Reganold provided feedback on earlier versions of this article. This work was funded by the Michigan State University Presidential Postdoctoral Fellowship in Ecology, Evolutionary and Behavior awarded to O.M.S.

Author contributions

All authors conceptualized the study. O.M.S. led the project. All authors collected data with greatest contributions from O.M.S. and C.L.D. O.M.S. analysed the data under guidance from K.L.D., W.L. and C.L.D. O.M.S., R.B.P., R.W., J.C.J., N.N. and C.L.D. made figures. O.M.S., R.B.P., B.F., L.N.J. and C.L.D. drafted paper sections with greatest contribution from O.M.S. O.M.S., K.L.D., R.B.P., R.W., K.C.D., B.F., J.C.J., L.N.J., W.L., N.N., E.E.C., A.H., C.M., A.N.S., O.J.U., M.L.Y. and C.L.D edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-023-01999-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-023-01999-w.

Correspondence and requests for materials should be addressed to Olivia M. Smith.

Peer review information *Nature Ecology & Evolution* thanks Sean Burns, Martin Nunez and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.