Voicing disagreement in science: Missing women*

David Klinowski[1]

February 2023

Abstract

This paper examines the authorship of post-publication criticisms in the scientific literature, with a focus on gender differences. Bibliometrics from journals in the natural and social sciences show that comments that criticize or correct a published study are 20-40% less likely than regular papers to have a female author. In preprints in the life sciences, prior to peer review, women are missing by 20-40% in failed replications compared to regular papers, but are not missing in successful replications. In an experiment, I then find large gender differences in willingness to point out and penalize a mistake in someone's work.

[1] Katz Graduate School of Business, University of Pittsburgh. dklinowski@katz.pitt.edu.

* This work was funded by Stanford University. For helpful comments and suggestions, I am grateful to Katherine Coffman, Josh Nicholson, Muriel Niederle, Collin Raymond, Alvin Roth, Colin Sullivan, Lise Vesterlund, and Alessandra Voena. This work also benefitted from feedback from participants in the Bay Area Behavioral and Experimental Economics Workshop, the Economic Science Association Conference at MIT Sloan, the University of East Anglia Discrimination and Disparities Workshop, the Maastricht Behavioral Economic Policy Symposium, the Stanford Institute for Theoretical Economics Conference in Experimental Economics, the Stanford Behavioral and Experimental Economics Workshop, the Economics of Gender Course at Stanford, the University of Pittsburgh Experimental Economics Workshop, and the Women and Public Policy Program Fellows Meetings at the Harvard Kennedy School.

1 Introduction

Debate and criticism of ideas are essential to science. By expressing disagreement and pointing out flaws in others' work, scientists refine their knowledge, correct the literature, and flag areas where consensus has not emerged. But, who voices disagreement in science? This paper investigates this question, with a focus on gender differences.

By many metrics, academic science remains male dominated. Women are a minority of all authors of scientific papers, and a minority of the faculty in most scientific fields (Handelsman et al., 2005; Ceci et al., 2014; Huang et al., 2020). While these empirical patterns are well studied, relatively little is known about women's participation specifically in papers that criticize published work. I this paper, I study women's authorship in two types of papers intended to express post-publication criticisms: comments and failed replications. I show that female authors are missing in these papers, across a range of scientific disciplines. This is true even after controlling for several possible explanations, including gender differences in sorting into fields, seniority, coauthorship, and priority for novelty.

Comments (i.e., post-publication critiques) are short papers or notes that express disagreement with, or point out flaws in, a published study. They are typically published in the same journal as that of the original paper. While relatively rare in economics, they are common in other fields, e.g., medicine.[1] As I will show, comments have some impact on the literature in terms of citations received and corrections issued in response. But comments can also take a confrontational tone.[2] Using bibliometric data, I study the authorship of comments in a set of high-

[1] The American Economic Review publishes 1 comment for every 17.5 regular papers, while the Journal of the American Medical Association publishes 1 comment for every 1.1 regular papers.

[2] In referring to comment-reply exchanges, Kahneman (2003) notes: "I [am] appalled by the absurdly competitive and adversarial nature of these exchanges, in which hardly anyone ever admits an error or acknowledges learning anything from the other." Journals have at times contributed to framing comments in a confrontational tone, as illustrated by the then official description of the comments section in Nature,

impact journals in the natural and social sciences. In these journals, the share of female authors hovers around 15-35% in regular papers, but trails at a 20-40% lower value in comments. Compared to regular papers, comments are authored by relatively junior scholars, which makes it all the more striking that so few women publish comments, given that women are relatively numerous among junior scholars. Moreover, female-solo-authored comments are exceptionally rare, across all journals examined. For example, in 21 years of data from the American Economic Review, one in every three comments with at least one male author is solo-authored, but there are no female-solo-authored comments. This lack of female-authored comments is not explained by gender differences in sorting into fields within journals, seniority, or coauthorship.

I then examine the authorship of replications that contradict previous findings, using bibliometrics from bioRxiv, the leading preprint repository in the life sciences. Papers on bioRxiv report either original work, successful replications, or failed replications, as indicated by the author(s) when they post their paper on the platform. The share of female authors in original work and successful replications hovers around 25-35% and is indistinguishable from one another. But it trails at a value 16-20% lower in failed replications, and 40% lower if I restrict the analysis to first authors. Thus, already before peer review, female authors are missing in failed replications but not in successful replications. This result suggests the reason women are missing in failed replications is unlikely to be that they prioritize novelty more so than men.

As I discuss in the next section, many factors may contribute to the observed gender gap in criticisms in the literature. I examine one candidate explanation with an experiment designed to test for gender differences in preferences for pointing out a mistake in someone's work and taking

---

one of the most prestigious and widely read scientific journals: "*Nature*'s Letters and Articles frequently stimulate responses from the authors' peers. In particular, a reader may submit an attack on the core of a paper. It is then our duty to take it up with the authors and referees. If the attack turns out to be well founded, a retraction or correction will follow." (Nature, 2004).

away credit earned from that mistake. In the experiment, a subject is paired with a participant who performs a task with some false positive rate (i.e., if the participant fails in the task, the task is marked as correct and awarded a payment with some probability). The subject must choose whether to take away the participant's payment and inform the participant that they failed in the task in the case of a false positive. While 40% of male subjects take up this option, only 16% of females do so. This gender gap in take-up is not explained by beliefs of own or others' performance on the task, and remains significant after accounting for distributional preferences. Thus, the experiment finds gender differences in preferences for pointing out a mistake in someone's work and deducting credit earned from it in an environment stripped of confounds present in academia.

That women do not participate in criticisms as much as men has implications for science as a labor market and for efforts to attract more women into science. How much is the process of knowledge accumulation and self-correction in science delayed by the lack of female-authored comments and failed replications? Are women's outcomes in academia, such as hiring and promotions, hurt by their (lack of) participation in criticisms? Would redesigning the incentives and institutions for criticisms increase women's participation in this part of the literature?

2 Theory

Many factors may influence a scholar's decision to express criticism of a published paper. Here, I consider a number of factors that have been identified in past work, and discuss how they might contribute to a gender gap in the authorship of comments and failed replications.

Seniority: In the sociology of science, Barber (1962) examines why scientific discoveries are sometimes met with resistance by other scientists. Barber argues that senior scholars are more likely than younger ones to resist new ideas, as their views are more easily colored by substantive

and methodological preconceptions and their higher status in the profession makes them more apt to criticize others' discoveries. Work in the economics of science also studies this question (Diamond, 1980; Stephan, 1996; Azoulay, Fons-Rosen, and Graff Zivin 2019). Accordingly, if male scholars are generally more senior than female scholars,[3] then gender differences in seniority might explain why so few comments are female-authored.

Professional returns: If comments and replications are regarded by the profession as lacking in value and originality (Camfield and Palmer-Jones, 2013; Duvendack, Palmer-Jones, and Reed, 2017), then it may be rational for career-minded scientists not to embark on these types of papers and focus instead on doing original work. Women particularly may prioritize novelty more so than men if they face greater time constraints and barriers to career advancement.

Behavioral factors: It is often unclear what implications comments and failed replications have for the original paper. At best they indicate an honest disagreement on the science; at worst they uncover fraud (Clemens, 2017). Whatever the case, criticisms may hurt the reputation of the original author and generate acrimonious debate (Kahneman, 2003; Camfield and Palmer-Jones, 2013; Galiani, Gertler, and Romero, 2017; Nosek and Errington, 2020). Thus, a scholar's decision to write a comment or a failed replication may be influenced by her preferences and beliefs, including her other-regarding concerns, her distaste for confrontation, her taste for uncovering and pointing out errors or fraud in others' work, her sense of the importance to science of correcting the literature, and her confidence in her own expertise on the subject (Boffey, 1988; Lacetera and Zirulia, 2011; Camfield and Palmer-Jones, 2013; Gelman, 2013). In some of these dimensions, researchers have identified gender differences. For example, Andreoni and Vesterlund (2001) and

[3] Nonnemaker (2000), Ceci et al. (2014), and Huang et al. (2020) document gender differences in seniority across several scientific disciplines, with women having on average fewer publications and shorter publication careers than men.

Klinowski (2018) identify gender differences in altruism. A large literature finds a gender gap in competitiveness, especially in male-typed domains (Niederle and Vesterlund, 2007; Healy and Pate, 2011; Dargnies, 2012; Shurchkov, 2012; Klinowski, 2019), which may be relevant if authors of criticisms "view themselves as on the attack" (Hamermesh, 2007). Several papers document that women are less likely than men to self-assess as competent, self-promote, and contribute their ideas, particularly in male-typed domains (Coffman, 2014; Karpowitz and Mendelberg, 2014; Coffman, Flikkema, and Shurchkov, 2021; Exley and Kessler, 2022; Murciano-Goroff, 2022). These gender differences could conceivably contribute to a gender gap in the authorship of criticisms.[4] [5]

Backlash: Work in psychology shows that women are penalized for displaying male stereotypical behavior such as assertiveness (Rudman and Glick, 2001; Heilman and Okimoto, 2007; Eagly and Carli, 2007; Williams and Tiedens, 2015). Women may forgo career-enhancing actions to avoid such penalties (Bursztyn, Fujiwara, and Pallais, 2017). Accordingly, if authors of comments and failed replications risk being perceived as assertive or confrontational, and this perception hurts women primarily, then women may avoid writing criticisms to avoid the backlash.[6]

[4] Relatedly, Collier and Bear (2012) find that women are less likely than men to contribute content on Wikipedia, in part because they dislike the level of conflict involved in editing others' entries. Wolak (2020) finds that men report greater enjoyment of arguments than women, which helps to explain their greater political engagement. Work in psychology finds that women use more polite language, less speaking time, and fewer interruptions that seek to dominate a conversation (Lakoff, 1975; Brown and Levinson, 1987; Jacklin and Maccoby, 1978; Holtgraves and Yang, 1992; Sheldon, 1993; Anderson and Leaper, 1998; Leaper and Ayres, 2007; Brescoll, 2011; Holmes, 2013). Relative to these studies, I provide evidence of gender gaps in participation in criticism in a labor market with incentives to participate.

[5] Relatedly, Shastry and Shurchkov (2021) find that young female economists respond more pessimistically than their male counterparts to a hypothetical journal rejection.

[6] Backlash may affect women's decisions to criticize others' work in several ways: (i) women may face more backlash than men, and accurately anticipate this; (ii) women may receive as much backlash as men but derive more disutility from it; (iii) women may anticipate more backlash than men, independent of true

The hypotheses above inform the analysis in the rest of the paper, where I seek to test or control for as many explanations as the data allow me to.[7] Finally, recent work has documented some of the findings I present: Wu et al. (2020) find a gender gap in the authorship of comments in PNAS and Science. Relative to Wu et al., I examine the authorship of comments in a larger set of journals and disciplines, look at failed replications in preprints in the life sciences, and investigate several explanations for the results using various analyses and an experiment.

3 Evidence from publications

3.1 Data

The data analyzed in this section consist of publications in the American Economic Review (AER), American Sociological Review (ASR), Journal of the American Medical Association (JAMA), Nature, Proceedings of the National Academy of Sciences (PNAS), and Science. These are high-impact journals that collectively cover a large set of natural and social sciences. These journals have published enough comments over time to make it possible to test for a gender gap in their authorship. I include in the dataset peer-reviewed research papers and comments, and exclude all other types of publications (e.g., editorials, non-peer-reviewed papers, notes about new work, and replies to comments).[8]

backlash and utility; and (iv) given a fixed perceived risk of backlash, women may be more averse to this risk than men. Thus, a backlash explanation may involve a behavioral mechanism.

[7] Figure A1 captures a Twitter discussion by scientists on the reasons (not) to pursue a replication "of a prominent scholar's work that has a flaw". Although only anecdotal, this discussion suggests scientists are influenced by factors described in Section 2, including seniority, risk of confrontation with other authors, professional returns, time pressure, other-regarding concerns, value to science, behavioral preferences such as "[being] OK at this because my job is, basically, to criticize the hell out of my peer-group's research", and backlash to women.

[8] The AER and ASR publish comments in a section called *Comments and Replies*, JAMA in *Comments and Responses* or *Letters to the Editor*, Nature in *Matters Arising*, PNAS in *Letters*, and Science in *Technical Comments*.

The period of analysis varies by journal and spans 7 to 22 years: it starts when the journal starts to publish comments or starts to distinguish comments from other papers in its online Table of Contents (eTOCs), and ends at the end of 2019. The Online Appendix describes the dataset construction, which included scraping eTOCs, assigning gender to authors using a commercial database and US Social Security Administration birth records, collecting citation data from Web of Science and Google Scholar, and constructing a proxy for author seniority from the author's cumulative number of publications. The sample consists of 488,499 article-author observations with gender assigned (Table A1).

3.2 Results

3.2.1 Share of female authors

Figure 1 plots the average share of female authors in regular papers and comments in the given time point, with observations at the article-author level with gender assigned. Trends are noisier for comments given their smaller sample size. Nevertheless, consistently across journals and most time points, the share of female authors is lower in comments than regular papers. The overall share of female authors in regular papers is 0.13 in the AER, 0.35 in the ASR, 0.34 in JAMA, 0.27 in Nature, 0.28 in PNAS, and 0.24 in Science. These values are in line with previous estimates of female authorship in economics, medicine, and other sciences (Jagsi et al., 2006; Ceci et al., 2014). In contrast, the share of female authors in comments is 0.08 in the AER, 0.24 in the ASR, 0.22 in JAMA, 0.22 in Nature, 0.20 in PNAS, and 0.16 in Science.

Table A2 presents results from ordinary least squares (OLS) regressions of the share of female authors in a journal on a comment indicator and year fixed effects. The estimated decline in female authorship in comments relative to regular papers is 40% in the AER (p=0.007), 32% in

the ASR (p=0.181), 37% in JAMA (p<0.001), 20% in Nature (p<0.001), 28% in PNAS (p<0.001), and 32% in Science (p<0.001). Results are robust to different methods of gender assignment and to analysis at the article level (Tables A3-A5).

3.2.2 Mechanisms

The fact that women are missing in comments across all journals examined suggests explanations that apply broadly across academia. This section examines several potential explanations.

Field sorting

Since comments are typically made by experts in the field of the original paper, gender differences in sorting into fields might explain the lack of female commenters. Specifically, if within a journal, women are relatively scarce in fields that naturally produce more comments, then women would mechanically be less represented in comments than regular papers. However, this explanation does not seem to account for the results. Table A6 shows estimates of the decline in female authorship in comments when controlling for field within journal, for the AER, Nature, and PNAS (these journals provide field information for their papers). After controlling for field, 81-89% of the gender gap in comments remains unexplained in the AER and PNAS, and the gap in fact widens slightly in Nature (see also Figure A2).

Seniority

If comments tend to be written by senior scholars, then gender differences in seniority might explain why so few comments are female-authored. I proxy for author seniority with the author's cumulative number of publications in the journal at the time of the observation. To better capture seniority, for AER authors I expand the dataset to include publications in all American

Economic Association journals in the period of analysis, and for JAMA and PNAS authors I include publications from further back in time than the main period of analysis (details in the Online Appendix, also Table A7). Several pieces of evidence indicate that the cumulative number of publications proxies well for seniority. First, authors listed last in a paper have significantly more cumulative publications than other authors, in journals for which the convention is to list the senior author last, but not in other journals (Table A8). Second, women have significantly fewer cumulative publications than men (Table A9), consistent with previous findings of greater male seniority in academia (Nonnemaker, 2000; Ceci et al., 2014; Huang et al., 2020). Lastly, the cumulative total number of publications is significantly larger for AER authors who also publish in the Journal of Economic Literature, a journal that features surveys of the economics literature often written by relatively experienced scholars (Table A10).

Table A9 shows that comment authors are in fact relatively junior. In all journals, comment authors have significantly fewer cumulative publications than authors of regular papers. This makes it even more striking that female authors are missing in comments, given that women are more represented among junior scholars. As a result, the estimated gender gap in comments widens after controlling for author seniority (Table A11). Controlling for author seniority, the estimated decline in female authorship in comments relative to regular papers is 45% in the AER ($p=0.003$), 33% in the ASR ($p=0.184$), 41% in JAMA ($p<0.001$), 24% in Nature ($p<0.001$), 36% in PNAS ($p<0.001$), and 33% in Science ($p<0.001$).

Coauthorship

If comments are frequently coauthored, perhaps women publish few comments because they coauthor less than men.[9] I investigate this possibility and find no support for it. Comments are in fact frequently solo-authored: the fraction of regular papers that are solo-authored vs. the fraction of comments that are solo-authored is 0.24 vs. 0.36 in the AER, 0.34 vs. 0.50 in the ASR, 0.02 vs. 0.32 in JAMA, 0.01 vs. 0.14 in Nature, 0.01 vs. 0.22 in PNAS, and 0.01 vs. 0.26 in Science (Figure A3).[10] And while it is true that women coauthor less than men in the AER and ASR, they coauthor more than men in JAMA, Nature, PNAS, and Science (Table A12a), which makes it difficult for gender differences in coauthorship to explain across the board the gender gap in commenting.

Notably, female-solo-authored comments are exceptionally rare. To illustrate this, Figure 2 plots the fraction of regular papers and comments that are solo-authored, conditional on the gender of at least one author in the article. The AER panel shows that of all regular papers in the AER with at least one male author, 19% are solo-authored (by a man, obviously), whereas of all comments with at least one male author, 33% are solo-authored. Thus, solo-authorship is more likely in comments than regular papers conditional on a male author. In contrast, conditional on a female author, 16% of regular papers are solo-authored, but no comment is female-solo-authored. All journals exhibit a similar pattern: for men, the propensity to solo-author increases considerably for comments relative to regular papers, while for women this propensity does not increase as much, and in fact decreases in the AER and ASR. This difference-in-difference in solo-authorship across paper type and gender is significant in all journals except the ASR (Table A12b). Together, these results indicate that lack of coauthors does not explain why women are missing in comments.

[9] In economics, Boschini and Sjögren (2007) and Hospido and Sanz (2019) find that women coauthor less than men, but Ductor, Goyal, and Prummer (2020) find opposite results.
[10] Note that JAMA requires comments to have at most three authors.

Time constraints

Three journals in the dataset require comments to be submitted within a specific period from the publication of the original paper: JAMA (4 weeks), Science (3 months), and PNAS (6 months). If these time limits are more binding for women, time constraints might explain why women publish relatively few comments. Time constraints might play a role even in journals with no time limits for submitting comments, if in practice comments in these journals are submitted quickly after the publication of the original paper. However, in the AER, the median comment is published 4.83 years after the publication of the original paper, and 10% of comments are published 11 or more years after the publication of the original paper. The same percentiles are 2.25 and 5.25 years in the ASR, and 0.86 and 2.82 years in Nature (Figure A3). Thus, many comments in the AER, ASR and Nature are published long after the publication of the original paper. The fact that women are just as underrepresented in comments in these journals as in journals with time limits suggests that time limits do not explain the gender gap in comments.

Returns and impact

Perhaps women write fewer criticisms than men because they prioritize novelty and impact more so than men. The analysis of successful vs. failed replications in the next section provides the strongest test I can give of this explanation. In this section, I give evidence of the impact comments can have on the author and the literature. One measure of impact is citations.[11] Table A13 and Figure A5 show how citations to comments compare to citations to regular papers in a journal. In the AER, regular papers in the 50th, 75th, 95th, and 99th percentiles of citations receive 5.7, 11.4, 27.9, and 61.2 annual citations, while comments in the same percentiles receive 1.1, 2.7, 8.5, and

[11] Citations are a measure of an article's influence on the literature, and are often used in promotion decisions (Lehmann, Jackson, and Lautrup, 2006; Ellison, 2013).

10.5 annual citations. Comments are clearly less cited than regular papers, but there is some overlap in the distributions. The 25% most cited comments in the AER receive more citations than 25% of all regular papers in the AER. Results are similar, but weaker, for other journals. In the ASR, Nature, PNAS, and Science, approximately 1-5% of comments receive more citations than the *median* regular paper in the same journal.[12] In contrast, most comments in JAMA receive few if any citations. And, while scientists tend to value original work over criticisms, publishing a comment in a top journal can be an opportunity for a scholar to "make his/her reputation" (Hamermesh, 2007).

Nature provides another measure of the impact comments can have on the literature. Any post-publication correction to a Nature paper is noted on Nature's website along with the correction date. From this information I compute how often Nature papers receive a comment and are subsequently retracted. Of the 232 Nature papers commented on between 2004 and 2019, 6 (2.6%) were retracted following the comment. By contrast, of the 13,039 Nature papers not commented on in the same period, 40 (0.3%) were retracted. This difference in retraction rates is suggestive that comments contribute to retractions. If so, then comments have not only a direct impact on the literature through their impact on retractions, but also a second-order impact given that retractions have been found to produce a decline in citations to the retracted paper and to other papers by the affected author, as well as a decline in new publications and funding flowing into the affected field (Furman, Jensen, and Murray, 2012; Azoulay et al., 2015; Azoulay, Bonatti, and Krieger, 2017).

[12] The overlap in the distributions of citations is not due to a large fraction of regular papers receiving no citations (Figure A6).

Papers commented on

Authors of papers commented on are relatively junior. They have fewer cumulative publications than authors of papers not commented on, both among all authors on a paper and in the sample restricted to the last author on a paper (Table A14). This result admits several interpretations, for example that papers by junior scholars are more likely to contain flaws, or that commenters shy away from criticizing senior authors, perhaps because of lack of self-confidence ("who am I to criticize this senior scientist?") or backlash ("I do not want to make an enemy out of this person"). As a crude, imperfect way to distinguish self-confidence from backlash, I estimate the likelihood that a paper in PNAS and Science receives a comment on an indicator that the last author on the paper is or has previously been an editor in the journal, conditional on that author's seniority. The rationale for this test is that conditional on the seniority of the last author on a paper, unwillingness to criticize the paper if the last author is an editor may reflect fear of backlash. I find some evidence of this mechanism in Science but not in PNAS. In Science, conditional on the seniority of the last author on the paper, the paper is significantly less likely to receive a comment if the last author is an editor (Table A15).[13]

4 Evidence from preprints

4.1 Data

The data analyzed in this section consist of all papers posted on bioRxiv, the leading preprint repository in the life sciences, between its launch in November 2013 and the end of 2019. Researchers use bioRxiv to disseminate their work prior to publication. Once posted on bioRxiv,

[13] Incidentally, authors on papers commented on are no more or less likely to be female than authors on other papers (Table A16). There is some evidence that women are more likely than men to comment on papers by other women in PNAS and Science (Table A17), and by junior authors in the AER (Table A18).

preprints can be edited but cannot be deleted. I take advantage of the fact that when posting a preprint, its author(s) must classify it as reporting either *New Results*, *Confirmatory Results*, or *Contradictory Results*. As defined by bioRxiv, "*New Results describe an advance in a field[,] Confirmatory Results largely replicate and confirm previously published work, whereas Contradictory Results largely replicate experimental approaches used in previously published work but the results contradict and/or do not support it*."

The sample consists of 351,662 paper-author observations with gender assigned, of which 98.3% are *New Results*, 1.3% are *Confirmatory Results*, and 0.5% are *Contradictory Results.* The Online Appendix describes the sample construction. Table B1 shows descriptive statistics.

4.2 Results

Figure 3a plots the average share of female authors in each paper type and time point, from observations at the paper-author level with gender assigned. The overall share of female authors is 0.32 in *New Results*, 0.32 in *Confirmatory Results*, and 0.26 in *Contradictory Results.* This is a 19% decline in female authorship in failed replications relative to other papers, significant at $p<0.001$ when estimated from OLS regressions, and robust to including year and field fixed effects (Table B2). The results are similar if estimated at the paper level (Table B3).

Figure 3b replicates Figure 3a on the sample restricted to the first author on each paper. In many life sciences, the convention is to list first the author who contributed most to the paper (Verhagens et al., 2003). Thus, any gender difference in motivations to publish a failed replication might be reflected more strongly among first authors. This is corroborated in Figure 3b. Among first authors, the overall share of female authors is 0.34 in *New Results*, 0.32 in *Confirmatory Results*, and 0.20 in *Contradictory Results*. This is a 38-41% decline in female authorship in failed replications relative to other papers, significant at $p<0.001$ when estimated from OLS regressions,

and robust to including field and year fixed effects (Table B2). Thus, the extent to which women are missing in failed replications is more than twice as large among first authors than among all authors on a paper.

Finally, women are also missing in failed replications in the sample restricted to solo-authored preprints. In this sample, the share of female authors is 10.4% in *New Results*, 13.8% in *Confirmatory Results*, and 5.9% in *Contradictory Results* (Figure B1). However, these values are not significantly different from each other, likely due to low power (Table B4).

In sum, across the life sciences, female authors are missing in failed replications, but not in successful replications. The reason women are missing in failed replications is thus unlikely to be that women prioritize novelty more so than men. Since preprints are posted on bioRxiv before peer review, one might conclude that peer review cannot explain the results. However, it is possible that women anticipate (correctly or not) unfavorable peer review to their failed replications, and, as a result, avoid writing these papers.[14]

5 Experimental evidence

As described in Section 2, several behavioral factors might contribute to a gender gap in post-publication criticisms. In this section, I investigate one candidate factor by running an experiment designed to test for gender differences in preferences for pointing out a mistake in someone's work and taking away credit earned from that mistake. I abstract away from other factors not to dismiss them as unimportant, but rather to examine whether gender differences in choices exist even absent such confounds.

[14] The evidence is mixed on whether women are discriminated against in general during peer review (in economics, see Blank, 1991; Abrevaya and Hamermesh, 2012; Card, et al., 2020; Hengel, 2020).

5.1 Experimental design

Task

The experiment involves a task of determining how many ones are in a matrix of 300 randomly generated zeros and ones (Abeler et al., 2011). Subjects have two minutes to solve the task, and receive $1 if they solve it correctly and $0 otherwise.

Choice to contest

At the beginning of the study, the participant—referred to here as "P1"—receives information on the task and completes an unincentivized two-minute practice round. Then, P1 is informed that he has been randomly matched to a participant from another session who performed the task, referred to here as “P2”. P2’s earnings from the task are noisily determined, with a 50% false positive rate. That is, P2 receives $1 if she solves the task correctly, and if she solves the task incorrectly she receives $0 with 50% chance and $1 with 50% chance. At the end of her session, P2 is informed of her earnings but not of whether she solved the task correctly. She is also informed that her final payment depends on the choice of a subject she is matched to, and that she will receive her payment after this choice is made. P2 makes no decision that affects P1.

After learning how P2's earnings are determined, P1 is asked whether he wishes to inform P2 that she solved the task incorrectly and deduct $1 from her earnings. Here, I refer to this choice as the choice to contest. If P1 chooses to contest, his choice is implemented only if (i) P2 received $1 by a false positive and (ii) P1 solves the task correctly. P1 makes his choice without knowing P2’s performance or earnings, and before he performs the task. If P1 chooses not to contest, or if (i) and (ii) do not hold, then P2 is paid her previously announced earnings and no additional information is sent to her. If P1 chooses to contest and (i) and (ii) hold, then P2 is paid her

previously announced earnings minus $1, and a message is sent to her informing her that she solved the task incorrectly and that the subject she was matched to chose to point this out and deduct $1. Regardless of his choice, P1 never receives feedback on P2's performance or earnings.

Eliciting P1's choice to contest as a decision contingent on (i) and (ii) minimizes the role P1's beliefs of performance might play in driving his choice. In general, a person's willingness to point out a mistake in someone's work may depend on how confident he is that there is a mistake and that he himself would not make the same mistake. In the experiment, these beliefs are irrelevant because P1's decision to contest is implemented only if P2 solves the task incorrectly and P1 solves it correctly.

Beliefs and performance

P1's beliefs of performance might nevertheless influence his choice to contest through indifference. Specifically, if P1 believes either that P2 solved the task correctly or that he will solve the task incorrectly, then P1 may be indifferent between contesting and not, since he believes contesting will not be implemented. To control for potential indifference in the analysis, I elicit P1's beliefs of his and P2's chances of solving the task correctly. After reporting these beliefs, P1 performs the task. Then, P1 performs a surprise additional round of the task, which serves to measure his performance without potential endogeneity of the choice (that is, in the first round of the task, P1 may be motivated to perform well because he wishes to implement his choice). Lastly, P1 completes a demographics questionnaire and the session ends.

Distributional preferences

Gender differences in the choice to contest might stem from gender differences in preferences over distributions of payoffs to P1 and P2. To examine this possibility, I conduct an additional treatment, administered to different subjects in an across-subject design, in which the subject (P1) is matched to a passive participant (P2). Rather than there being a task and a choice to contest, P1 simply makes a choice over two matrices that describe payoffs to P1 and P2 (Table C1). The matrices are designed to mirror the potential payoffs P1 and P2 receive in the main treatment given P1's choice to contest. Just as choosing to contest reduces P2's payoff by $1 if (i) and (ii) occur, the only difference between the two matrices is that one of them reduces P2's payoff by $1 in a specific contingency while the other matrix does not reduce P2's payoff. The probability that the contingency occurs is exogenous and calibrated to match typical beliefs P1 holds in the main treatment about the chance that (i) and (ii) occur. Thus, observed choices over matrices help to benchmark the role distributional preferences play in driving observed choices to contest in the main treatment. Hereafter, I refer to the main treatment as the *Contest treatment* and the additional treatment as the *Deduct-a-$1 treatment*. The sample comprises 301 subjects in the *Contest treatment* and 188 subjects in the *Deduct-a-$1 treatment* (Table C2 for descriptive statistics). The Online Appendix describes the implementation of the experiment. The supplementary files include the experimental instructions.[15]

[15] Relatedly, experiments by Isaksson (2018) and Guo and Recalde (2022) investigate individuals' willingness to correct the move of a partner in a group task.

5.2 Results

Figure 4 plots the fraction of subjects who choose to contest in the *Contest treatment:* 40% of men and 16% of women ($p<0.001$). The estimated gender difference is 19.1 percentage points ($p<0.001$) in a regression that controls for performance on the task, beliefs of performance, potential indifference in the choice, and demographics (Table C3). Figure 4 also plots the fraction of subjects who choose to deduct the $1 in the *Deduct-a-$1 treatment:* 16% of men and 4% of women ($p=0.009$). The estimated gender difference in this latter choice is 10.8 percentage points ($p<0.001$) in a regression that controls for demographics (Table C3). The estimated difference-in-differences in choices across gender is 12.2 percentage points ($p=0.066$) without demographic controls and 11.5 percentage points ($p=0.084$) with demographic controls (Table C3).

Thus, women are 50% less likely than men to choose to inform P2 that she solved the task incorrectly and reduce her payoff. Half of this gender gap is accounted for by distributional preferences. The residual half of the gender gap is significant, which indicates there exist gender differences in payoff-irrelevant preferences for pointing out the mistake on the task and deducting credit earned from it.[16]

6 Discussion

This paper documents that the standard mechanisms journals use to communicate post-publication criticisms—comments and failed replications—are dominated by male authors. What can be done to increase female participation in criticisms in science? One approach might be to exhort women to write criticisms, but research suggests this approach may not work and may backfire if writing

[16] I conducted additional treatments that introduce a monetary incentive for P1 to contest. In these treatments, men continue to choose to contest significantly more often than women (Online Appendix).

comments and failed replications brings disutility or backlash to women, as Exley, Niederle, and Vesterlund, (2020) note regarding calls for women to "lean in" in response to the gender wage gap. A more fruitful approach may be the market design approach (Roth, 2018), which in this case calls for rethinking the mechanisms through which scientists voice criticisms and the incentives they have to do so. One obvious start in this direction may be for journals not to frame criticisms as attacks, and instead emphasize their contribution to knowledge production. Future research may examine whether proposals intended to increase replications, such as for journals to commission replications with guaranteed publication (Hamermesh, 2007) or create replications sections similar to existing comments sections (Coffman, Niederle, and Wilson, 2017), would attract men and women equally. Given the central role criticisms play in science, careful design of the mechanisms and incentives for expressing criticisms may help to attract more women into academic science and improve the reliability of the scientific literature.

REFERENCES

Abrevaya, J., and Hamermesh, D.S., 2012. Charity and favoritism in the field: Are female economists nicer (to each other)? *Review of Economic and Statistics*, 94.1: 202-207.

Anderson, K.J., and Leaper, C., 1998. Meta-analyses of gender effects on conversational interruption: Who, what, when, where, and how. *Sex Roles*, 39.3/4: 225-252.

Andreoni, J., and Vesterlund, 2001. Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics*, 116.1: 293-312.

Azoulay, P., Bonatti, A., and Krieger, J.L., 2017. The career effects of scandal: Evidence from scientific retractions. *Research Policy*, 46.9: 1552-1569.

Azoulay, P., Fons-Rosen, C., and Graff-Zivin, J.S., 2019. Does science advance one funeral at a time? *American Economic Review*, 109.8: 2889-2920.

Azoulay, P., Furman, J.L., Krieger, J.L., and Murray, F.E., 2015. Retractions. *Review of Economics and Statistics*, 97.5: 1118-1136.

Barber, B., 1961. Resistance by scientists to scientific discovery. *Science*, 134.3479: 569-602.

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

Boffey, P.M., 1988. Two critics of science revel in the role. *The New York Times*, April 18.

Boschini, A., and Sjögren, A., 2007. Is team formation gender neutral? Evidence from coauthorship patterns. *Journal of Labor Economics*, 25.2: 325-365.

Brescoll, V.L., 2011. Who takes the floor and why: Gender, power, and volubility in organizations. *Administrative Science Quarterly*, 56.4: 622-641.

Brown, P., and Levinson, S.C., 1987. *Politeness: Some Universals in Language Usage.* Vol. 4, Cambridge University Press.

Bursztyn, L., Fujiwara, T., and Pallais, A., 2017. 'Acting wife': Marriage market incentives and labor market investments. *American Economic Review*, 107.11: 3288-3319.

Camfield, L., and Palmer-Jones, R., 2013. Three `Rs' of econometrics: Repetition, reproduction, and replication. *Journal of Development Economics*, 49.12: 1607-1614.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N., 2020. Are referees and editors in economics gender neutral? *Quarterly Journal of Economics*, 135.1: 269-327.

Ceci, S.J., Ginther, D.K., Kahn, S., and Williams, W.M., 2014. Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15.3: 75-141.

Clemens, M.A., 2017. The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, 31.1: 326-342.

Coffman, K.B., 2014. Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics*, 129.4: 1625-1660.

Coffman, K.B., Flikkema, C.B., and Shurchkov, O., 2021. Gender stereotypes in deliberation and team decisions. *Games and Economic Behavior*, 129: 329-349.

Coffman, L.C., Niederle, M., and Wilson, A.J., 2017. A proposal to organize and promote replications. *American Economic Review: Papers & Proceedings*, 107.5: 41-45.

Collier, B., and Bear, J., 2012. Conflict, confidence, or criticism: An empirical examination of the gender gap in Wikipedia. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 383-392.

Dargnies, M.-P., 2012. Men too can sometimes shy away from competition: The case of team competition. *Management Science*, 58.11: 1982-2000.

Diamond, A.M., 1980. Age and the acceptance of cliometrics. *Journal of Economic History*, 40.4: 838-841.

Ductor, L., Goyal, S., and Prummer, A., 2020. Gender and collaboration. *Working paper*.

Duvendack, M., Palmer-Jones, R., and Reed, W.R., 2017. What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review*, 107.5: 46-51.

Eagly, A.H., and Carli, L.L., 2007. *Through the Labyrinth: The Truth about How Women Become Leaders*. Harvard University Press.

Ellison, G., 2013. How does the market use citation data? The Hirsch index in economics. *American Economic Journal: Applied Economics*, 5.3: 63-90.

Exley, C.L., Niederle, M., and Vesterlund, L., 2020. Knowing when to ask: The cost of leaning in. *Journal of Political Economy*, 128.3: 816-854.

Furman, J.L., Jensen, K., and Murray, F., 2012. Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy*, 41.2: 276-290.

Galiani, S., Gertler, P., and Romero, M., 2017. Incentives for replication in economics. NBER working paper 23576.

Gelman, A., 2013. Ethics and statistics: It's too hard to publish criticisms and obtain data for replication. *Chance*, 26.3: 49-52.

Guo, J., and Recalde, M.P., 2022. Overriding in teams: The role of beliefs, social image, and gender. *Management Science*, forthcoming.

Hamermesh, D.S., 2007. Replication in economics. *Canadian Journal of Economics*, 40.3: 715-733.

Handelsman, J., Cantor, N., Carnes, M., Denton, D., Fine, E., Grosz, B., Hinshaw, V., Marrett, C., Rosser, S., Shalala, D., and Sheridan, J., 2005. More women in science. *Science*, 309.5738: 1190-1191.

Healy, A., and Pate, J., 2011. Can teams help to close the gender competition gap? *Economic Journal*, 121.555: 1192-1204.

Heilman, M.E., and Okimoto, T.G., 2007. Why are women penalized for success at male tasks? The implied communality deficit. *Journal of Applied Psychology*, 92.1: 81-92.

Hengel, E., 2020. Publishing while female. *Working paper*.

Holmes, J., 2013. *Women, Men, and Politeness.* Routledge.

Holtgraves, T., and Yang, J.N., 1992. Interpersonal underpinnings of request strategies: General principles and differences due to culture and gender. *Journal of Personality and Social Psychology*, 62.2: 246-256.

Hospido, L., and Sanz, C., 2019. Gender gaps in the evaluation of research: Evidence from submissions to economics conferences. *IZA Discussion Paper 12494.*

Huang, J., Gates, A.J., Sinatra, R., and Barabási, A.L., 2020. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, 117.9: 4609-4616.

Isaksson, S., 2018. It takes two: Gender differences in group work. Working paper.

Jacklin, C.N., and Maccoby, E.E., 1978. Social behavior at thirty-three months in same-sex and mixed-sex dyads. *Child Development*, 49.3: 557-569.

Jagsi, R., Guancial, E.A., Worobey, C.C., Henault, L.E., Chang, Y., Starr, R., Tarbell, N.J., and Hylek, E.M., 2006. The "gender gap" in authorship of academic medical literature—A 35-year perspective. *New England Journal of Medicine*, 355.3: 281-287.

Kahneman, D., 2003. Experiences in collaborative research. *American Psychologist*, 58.9: 723-730.

Karpowitz, C.F., and Mendelberg, T., 2014. *The Silent Sex: Gender, Deliberation, and Institutions*. Princeton University Press.

Klinowski, D., 2018. Gender differences in giving in the Dictator Game: The role of reluctant altruism. *Journal of the Economic Science Association*, 4.2: 110-122.

Klinowski, D., 2019. Selection into self-improvement and competition pay: Gender, stereotypes, and earnings volatility. *Journal of Economic Behavior & Organization*, 158: 128-146.

Lacetera, N., and Zirulia, L., 2011. The economics of scientific misconduct. *The Journal of Law, Economics, & Organization*, 27.3: 568-603.

Lakoff, R.T., 1975. *Language and Woman's Place*. Harper & Row.

Leaper, C., and Ayres, M.M., 2007. A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *Personality and Social Psychology*, 11.4: 328-363.

Lehmann, S., Jackson, A.D., and Lautrup, B.E., 2006. Measures for measures. *Nature*, 444.7122: 1003-1004.

Murciano-Goroff, R., 2022. Missing women in tech: The labor market for highly skilled software engineers. *Management Science*, 68.5: 3262-3281.

Nature, 2004. Enhancing *Nature*'s services. *Nature*, 428.6979: 105.

Niederle, M., and Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, 122.3: 1067-1101.

Nonnemaker, L., 2000. Women physicians in academic medicine—New insights from cohort studies. *New England Journal of Medicine*, 342: 399-405.

Nosek, B.A., and Errington, T.M., 2020. The best time to argue about what a replication means? Before you do it. *Nature*, 583: 18-520.

Roth, A.E., 2018. Marketplaces, markets, and market design. *American Economic Review*, 108.7: 1609-1658.

Rudman, L.A., and Glick, P., 2001. Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, 57.4: 743-762.

Shastry, G.K., and Shurchkov, O., 2021. Reject or revise: Gender differences in persistence and publishing in economics. Working paper.

Sheldon, A., 1993. Pickle fights: Gendered talk in preschool disputes. *Gender and Conversational Interaction*, 13.1: 83-109.

Shurchkov, O., 2012. Under pressure: Gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, 10.5: 1189-1213.

Stephan, P.E., 1996. The economics of science. *Journal of Economic Literature*, 34.3: 1199-1235.

Verhagens, J.V., Wallace, K.J., Collins, S.C., and Thomas, T.R., 2003. QUAD system offers fair shares to all authors. *Nature*, 426: 602.

Williams, M.J., and Tiedens, L.Z., 2016. The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior. *Psychological Bulletin*, 142.2: 165-197.

Wolak, J. 2020. Conflict avoidance and gender gaps in political engagement. *Political Behavior*, 1-24.

Wu, C., Fuller, S., Shi, Z., and Wilkes, R., 2020. The gender gap in commenting: Women are less likely than men to comment on (men's) published research. *PLoS One*, 15.4: e0230043.

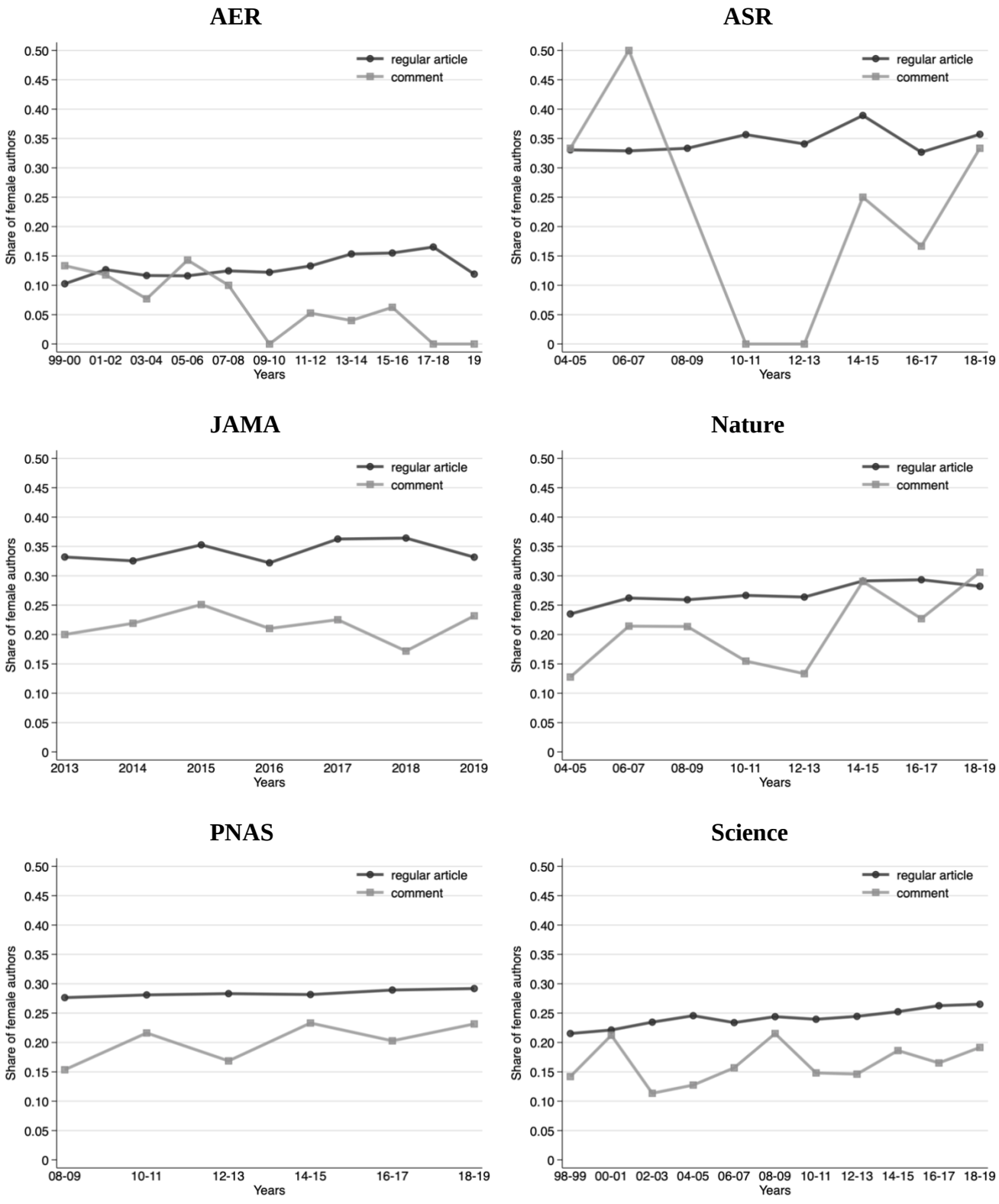

Figure 1 Share of female authors

Notes: Values are averages in the given time point from observations at the article-author level.

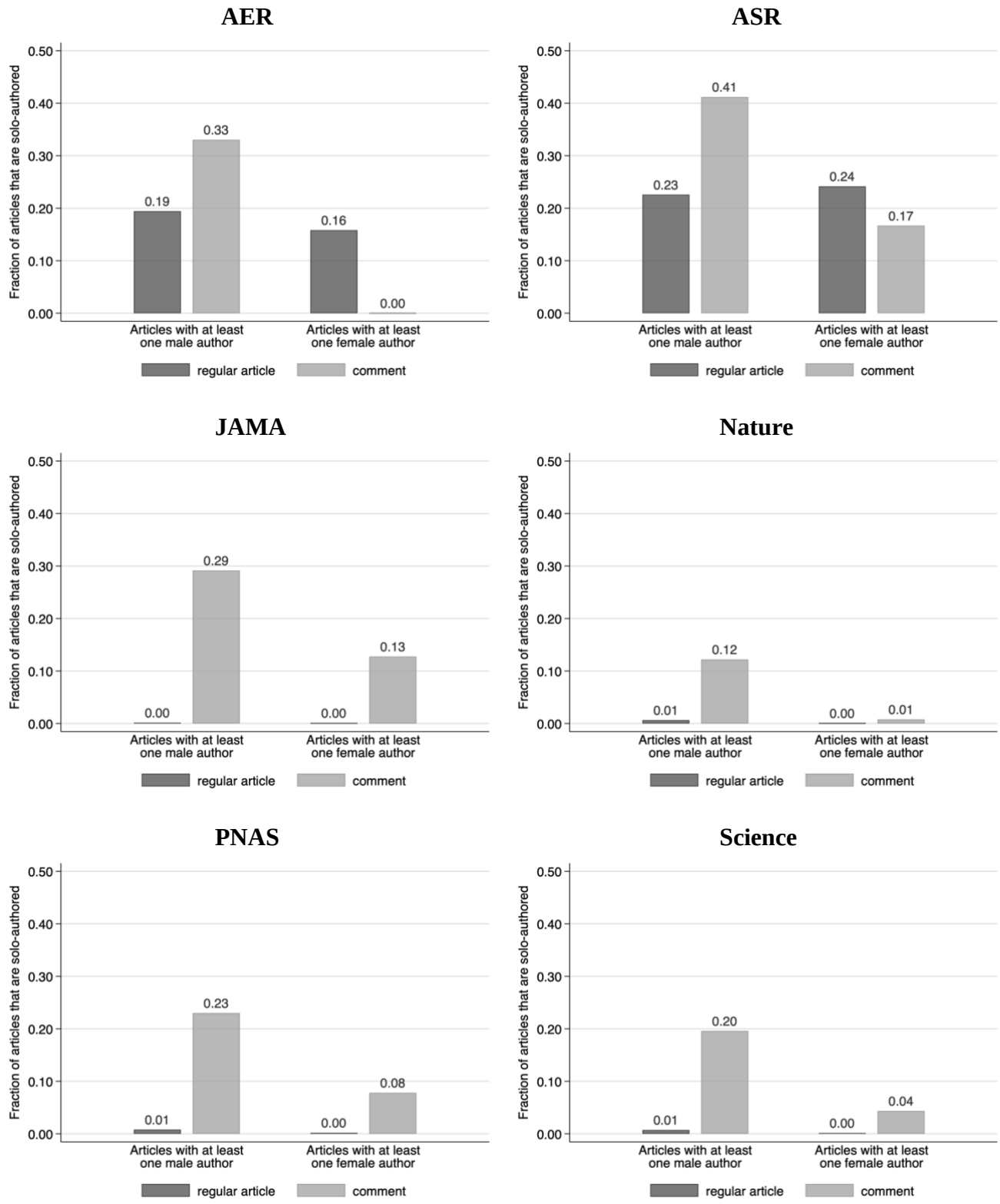

Figure 2 Fraction of articles that are solo-authored, conditional on the gender of at least one author
Notes: Observations at the article level. Articles with at least one male author and articles with at least one female author are not mutually exclusive; thus, the conditional samples across gender are not disjoint.

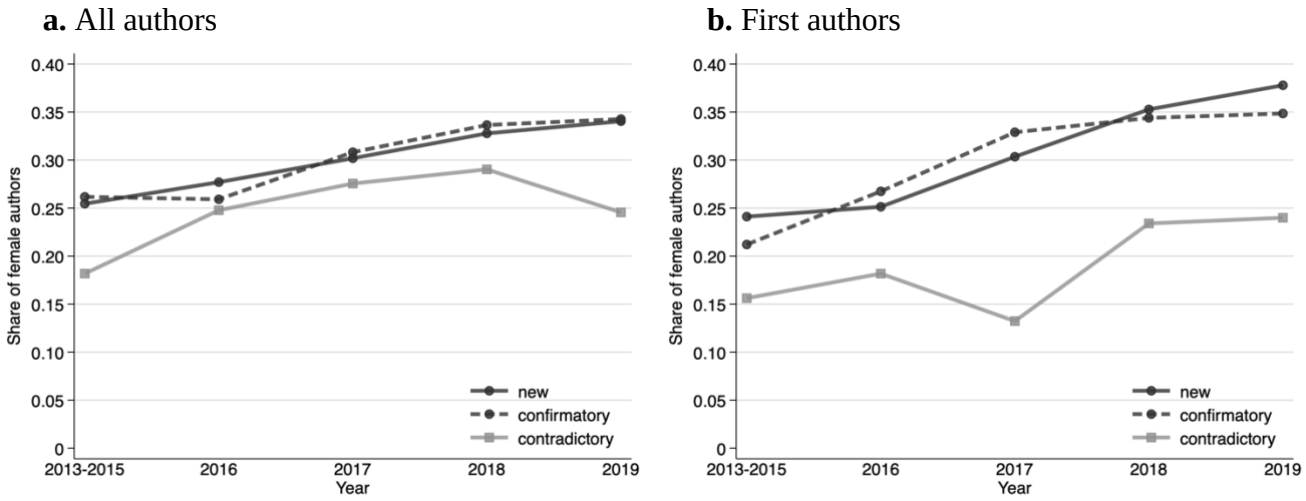

Figure 3 Share of female authors in bioRxiv preprints

Notes: Observations at the paper-author level, pooled across 2013-2015 due to the small sample in this period. A total of 107 papers were posted on bioRxiv in 2013; 853 in 2014; 1,793 in 2015; 4,750 in 2016; 11,460 in 2017; 20,780 in 2018; and 29,232 in 2019.

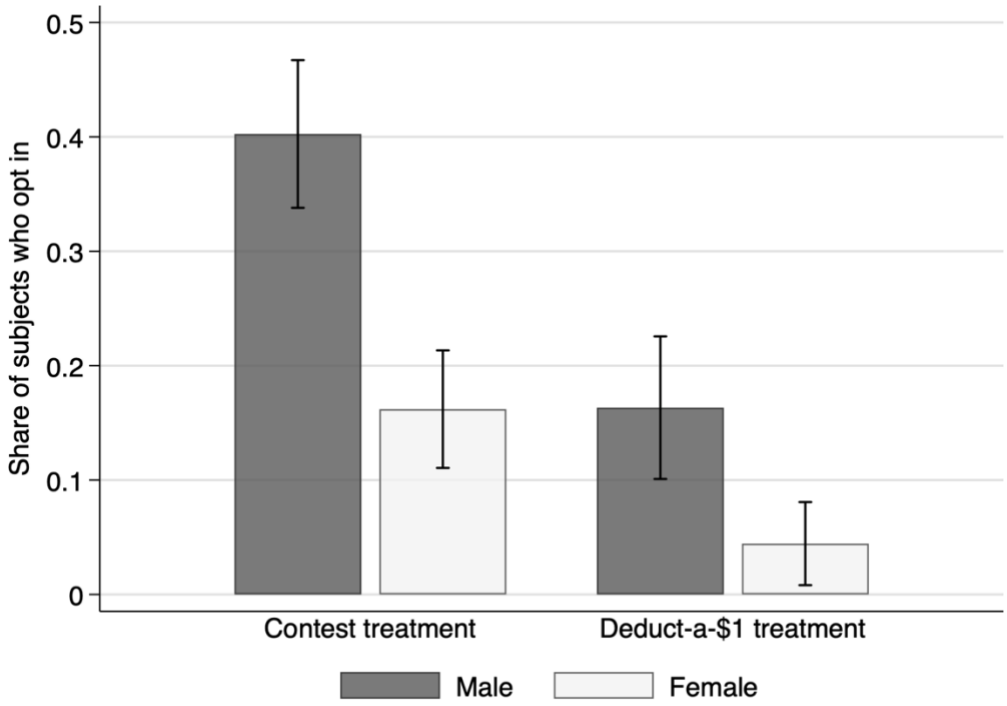

Figure 4 Take-up of contest option and equivalent payoff matrix in the experiment

Notes: Fraction of subjects who choose to contest in the *Contest treatment* and choose to deduct the $1 in the *Deduct-a-$1 treatment*. Whiskers indicate 90-percent confidence intervals.