Are Referees and Editors in Economics Gender Neutral?*

David Card
UC Berkeley and NBER

Stefano DellaVigna
UC Berkeley and NBER

Patricia Funk
Università della Svizzera Italiana

Nagore Iriberri
University of the Basque Country and Ikerbasque

September 2019

Abstract

We study the role of gender in the evaluation of economic research using submissions to four leading journals. We find that referee gender has no effect on the relative assessment of female- versus male-authored papers, suggesting that any *differential* biases of male referees are negligible. To determine whether referees as a whole impose different standards for female authors, we compare citations for female and male-authored papers, holding constant referee evaluations and other characteristics. We find that female-authored papers receive about 25% more citations than observably similar male-authored papers. Editors largely follow the referees, resulting in a 1.7 percentage point lower probability of a revise and resubmit verdict for papers with female authors relative to a citation-maximizing benchmark. In their desk rejection decisions, editors treat female authors more favorably, though they still impose a higher bar than would be implied by citation-maximization. We find no differences in the informativeness of female versus male referees, or in the weight that editors place on the recommendations of female versus male referees. We also find no differences in editorial delays for female versus male-authored papers.

*We thank the editors and staff of the *Journal of the European Economics Association*, the *Quarterly Journal of Economics*, the *Review of Economics and Statistics*, and the *Review of Economic Studies* for their invaluable assistance and support. We also thank Dan Hamermesh, Lawrence Katz, Andrei Shleifer, Lise Versterlund, and Heidi Williams for helpful comments, and Manu Garcia for sharing code. We are also grateful to Luisa Cefala', Margaret Chen, Alden Cheng, Manu Garcia, Johannes Hermle, Giovanni Kraushaar, Christopher Lim, Andrew Tai, and a team of undergraduate research assistants for their extraordinary help. Nagore Iriberri acknowledges financial support from grants ECO2015-66027-P MINECO/FEDER and IT869-13. Patricia Funk acknowledges financial support from the Swiss National Science Foundation (grant 178887). The analysis plan is posted on the AEA registry under AEARCTR-0003048.

1 Introduction

Women are under-represented in the top ranks of many professions, including corporate management (Bertrand and Hallock, 2001), law (Azmat and Ferrer, 2017), and academia (Ceci et al., 2014). While numerous explanations have been offered for this gap, including differences in competitiveness (Niederle and Vesterlund, 2011; Reuben et al., 2015) and in the allocation of time between work and family (Goldin, 2014), an abiding concern is that stereotype biases (Reuben et al., 2014; Bordalo et al., 2019; Bohren et al., forthcoming) or other forms of discrimination lead decision makers to *undervalue* the contributions of women. This concern is particularly salient in economics, where the vast majority of gatekeepers – senior faculty, journal editors, and referees – are male (Ginther and Kahn, 2004; Bayer and Rouse, 2016; Lundberg, 2017).

Existing evidence on gender biases in the evaluation of economic research is mixed. Blank (1991) randomly assigned submissions at the *American Economic Review* to referees, with or without masking the author's name and affiliation. She found a large but imprecisely estimated positive effect of blinding on the acceptance rate of female-authored papers. Broder (1993) analyzed reviews of NSF proposals, finding that female reviewers give lower average ratings to female-authored proposals. Abrevaya and Hamermesh (2012) find no significant gap in the evaluations of male- versus female-authored papers by male and female referees at an anonymous economics journal. Chari and Goldsmith-Pinkham (2017) likewise find no disparity in the acceptance rates of female- and male-authored papers for NBER conferences, though Hospido and and Sanz (2019) find a significant advantage for male authors at three different European conferences. Similarly, Hengel (2018) finds that female authors face a higher bar in the journal review process. Focusing on the general climate in the field, Wu (forthcoming) shows that online discussions often contain derogatory personal comments about female economists. Nevertheless, Donald and Hamermesh (2006) conclude that members of the AEA exhibit a *positive* preference for female candidates for the Association's executive board.

In this paper we analyze the role of gender in the evaluation process, using anonymized data on nearly 30,000 submissions to four leading economics journals: the *Journal of the European Economics Association*, the *Quarterly Journal of Economics*, the *Review of Economics and Statistics*, and the *Review of Economic Studies.* We combine paper characteristics – including the gender and previous publication record of the authors – with information on the gender of the referees assigned to the paper, their summary recommendations, the editor's decision, and the ultimate citations received by the paper, regardless of whether it was accepted or not.[1] We use these data to analyze gender differences in how papers are assigned to referees, how they are reviewed, and how editors use referee inputs to reach a revise and resubmit (R&R) verdict.[2] We also examine the

[1]We do not have access to any textual information in the referee report, the editorial letter, or the paper itself. Our access agreement required us to extract the data from each journal's archive using a stand-alone program that created a file with limited information on each paper, and no identifying information on authors or referees.

[2]Since the vast majority of editors in our 2003-2013 sample period are male it is not possible to examine the impact of editor gender. The evidence in Bransch and Kvasnick (2017) suggests that having female editors does not increase the share of female-authored papers published in top journals.

effects of author gender on desk-rejection decisions and editorial delays. Our analysis follows the analysis plan AEARCTR-0003048, which we drafted prior to the completion of our data collection to address concerns over data mining (see Christensen and Miguel, 2018).

We complement this data set with a survey of 141 economists, allowing us to compare our findings with the expectations of practitioners, as in DellaVigna and Pope (2018). We also elicit beliefs about gender differences in the citation-quality link to aid in interpreting our findings.

We begin in Section 2 with an overview of our submissions database. This builds on the 4-journal sample collected by Card and DellaVigna (forthcoming), hereafter CDV, adding information on the gender of authors and referees and the publication record of individual co-authors. We obtained the names of authors and referees from each journal prior to our data extraction, then used a combination of first-name coding and internet search to assign genders.[3] Two-thirds of the submissions were written by all-male teams of authors, 8% by all-female teams, and 19% by mixed-gender teams. We distinguish mixed-gender teams by whether the most-published ("senior") co-author is female (3% of all submissions) or not (16%), yielding four gender-mix categories.

In Section 3 we analyze the matching process by which papers are assigned to referees. As in earlier studies (e.g., Dolado et al., 2012; Lundberg, 2017; Chari and Goldsmith-Pinkham, 2017), we find that the share of female authors varies widely across fields, with roughly proportional variation in the share of female referees. Even controlling for field and other factors, however, editors are 7 percentage points (50%) more likely to assign a female-authored paper to a female referee, suggesting that they are sensitive to gender-related issues in the review process.

In Section 4, we conduct a simple audit-style comparison of the summary evaluations submitted by female and male referees. Our most general models include paper fixed effects, allowing us to quantify differences in assessments of *the same paper* by referees of each gender. Consistent with the findings of Abrevaya and Hamermesh (2012) we find that the relative evaluations of female and male-authored papers by male and female referees are very similar.

Although these findings rule out any *relative* bias in males' assessments of female-authored papers, they do not allow us to determine whether male and female authors face the same standards for publishing their work. To make further progress we need to make *between-paper* comparisons, accounting for differences in the quality of female- and male-authored papers. While there is no perfect measure of quality, we observe citations, which are correlated with quality and are highly relevant to publishers and editors. We therefore use citations as a noisy measure of quality, taking account of potential sources of divergence between the two.

We frame our analysis using the simple model developed in CDV. In this model, referee preferences for papers by different author groups can be inferred from the relationship between citations, referee recommendations, and author characteristics. If referees are unbiased judges, papers by male and female authors that receive similar referee evaluations will receive similar citations. Likewise, editors' preferences can be inferred by comparing the relative effects of author characteristics and

[3] We validate our approach using a sample of published papers in the same journals, showing that we can assign gender to the authors of 97% of these papers with misclassification rates of less than 1 percent.

referee recommendations on citations and R&R probabilities. If editors select papers to maximize expected citations, the relative weights for the referee recommendations and author characteristics in the R&R decision model will be proportional to their relative weights in the citation model.

There are two important caveats. First, if published papers receive more citations, conditional on quality, there is a potential "publication bias" in the relationship between paper characteristics and realized citations. As in CDV, we address this by including in the citation models an indicator for R&R status and a control function that corrects for endogeneity in the editor's decision.

Second, there are potential gender biases in citations. Indeed, our survey respondents believe that females receive about 6% fewer citations than males, holding constant the quality of their papers. To the extent that males receive more citations, conditional on quality, a finding of equal citations for male- and female-authored papers with similar referee recommendations means that the referees are actually setting a higher bar for female-authored papers. In this case, we obtain a bound on the gender gap in the referee recommendations.

We find that female-authored papers receive 22 log points (s.e.=0.05) *more* citations than male-authored papers, controlling for the referee evaluations. Our estimate of this gender gap is robust to alternative measures of citations and to a variety of alternative specifications. The magnitude of the gap *does* depend on whether we control for the prior publications of authors, since females have fewer prior publications and prior publications strongly predict citations. It also falls slightly (to 17 log points, with s.e.=0.05) if we add controls for the institutional affiliation of authors.[4]

For mixed-gender papers with a senior male co-author we find no difference in citations relative to those authored by all-male teams, consistent with our survey respondents' expectations that such papers are treated about the same as male-authored papers. For mixed-gender papers with a senior female co-author, however, we find a 6 log point (s.e.=0.07) citation premium. We cannot reject that this premium is one-half as large as the premium for papers written by all-female teams, again consistent with our survey respondents' expectations about how such papers are treated.

In Section 5 we study the R&R decisions of editors, given the referee recommendations and the characteristics of authors and their papers. On average editors tend to follow the referees' recommendations, putting essentially no weight on author gender in their R&R decisions. This means that they are *over-rejecting* female-authored papers relative to a citation-maximizing benchmark.

We then use our framework to compute how the R&R rate would change if editors were to reset the gender effects in their decision model to maximize expected citations. We estimate that the R&R rate for NDR papers with at least one female author—that is, averaging across all-female author teams and mixed gender teams—would increase from 14.2% to 15.9%, a sizable 12 percent increase. We find a similar impact in a second counterfactual in which editors assign the citation-maximizing weight not just to the gender variables, but also to the author publication variables.

There are two main explanations for our finding that female-authored papers receive more citations, conditional on the referee evaluations. The first is that referees hold female authors to a

[4]We use as our benchmark specification the one without institutional prominence since it was the one pre-specified in the analysis plan, but we consider a large number of alternative specifications in our robustness tables.

higher bar, perhaps because of stereotype biases. The second is that female-authored papers have characteristics that lead to higher citations but are not as highly rewarded in the review process. For example, female authors may tend to write more empirically-oriented papers, or concentrate on certain topics within broad field categories that referees undervalue relative to expected citations.[5]

To provide some evidence on this second explanation, we conduct an in-depth analysis of 1,719 papers published in the four journals in our sample in 2008-15 (approximately the period that accepted papers in our main sample would be published). In this sample we find a citation premium of 12 log points (s.e.=0.13) for female-authored papers and 24 log points (s.e.=0.13) for mixed-gender papers with senior female authors, controlling for the same covariates as in our main analysis, apart from the referee evaluations. These premiums are much less precisely estimated than those from our main sample (which is nearly ten times larger), confirming the advantages of our main approach. Nonetheless, they replicate the key patterns, and are qualitatively consistent with the findings of Hengel (2019) for a different sample of published papers.

We then add controls that are unavailable in our main sample. First, we include detailed 2-digit JEL field controls. These controls nearly double the R^2 in the citation regression, but barely affect the gender coefficients. Second, we include a series of controls to capture the theoretical versus empirical content of a paper based on word counts (words such as "proposition" and "standard error") and on qualitative ratings. Papers with more formal modeling receive fewer citations, even controlling for field, and female-authored papers tend to have less formal modeling. Thus, controlling for measures of content leads to somewhat smaller citation premiums, 9 log points (s.e.=0.13) for female-authored papers versus a baseline of 12 log points, and 13 log points (s.e.=0.15) for mixed-gender papers with senior female authors versus a baseline of 24 points. This provides suggestive evidence that paper characteristics play some role in explaining the gender gap in citations, though given the imprecision of the estimated premiums we cannot reach a definitive conclusion.

Returning to our main sample of submissions, we examine the desk-rejection process, which provides information on editors' preferences free from any referee inputs. Across all submissions (refereed or not), female-authored papers receive 24 log points more citations than male-authored papers, conditional on other controls. An editor who sets a bar for desk rejection based on expected citations should therefore give positive weight to female authorship. Consistent with this prediction, we find that editors desk reject fewer female-authored papers, holding constant other characteristics. The gap, however, is smaller than predicted by a citation-maximizing benchmark.

We then address two further issues in the editorial process: Are some referees more reliable in judging quality (as revealed by citations)? Do editors pay more attention to more reliable referees? CDV find that the recommendations of more and less prolific referees are equally predictive of future citations, yet editors tend to place more weight on recommendations from referees with more prior publications. In the case of gender, we find that male and female referees are equally informative, and that editors place equal weight on their recommendations in reaching an R&R decision. Thus,

[5] A third explanation, that this gap reflects a tendency for female authors to submit papers that have been circulating longer, is unlikely given that we find similar findings using the SSCI citations, which accrue only to published papers.

editors appear to be gender-neutral in their use of referee inputs.

Finally, in Section 6 we consider the speed of the review process. We find no gender differences in the time that referees take to return a recommendation, in the time that editors take to reach a decision, or in the time between submission and acceptance for published papers.

In light of these results, in our concluding section we offer a partial reconciliation of the myriad findings in the literature. Some of these differences appear due, at least in part, to the different strategies used to identify potential gender discrimination. One strategy, used by Broder (1993) and Abrevaya and Hamermesh (2012), is an audit-style comparison of the recommendations provided by different reviewers of male- and female-authored papers. These studies, like us, find no evidence of differential biases against female-authored papers by male referees. This suggests that the animus documented by Wu (forthcoming) against female economists in a prominent online discussion board is absent in the review process.

A second strategy is to compare outcomes for male- and female-authored papers without explicit quality controls, e.g, the analysis of NBER submissions by Chari and Goldsmith-Pinkham (2017). Like these authors, we find that female- and male-authored papers have similar R&R rates when we do not control for prior author publications. A third strategy is to compare outcomes conditional on quality controls. We are aware of only one prior analysis using this design – Donald and Hamermesh (2006) – which comes to the opposite conclusion as us, albeit in a different setting (the election of AEA officers). Finally, a fourth strategy is to compare outcomes when author identity (including gender) is blinded. The only such study we are aware of, by Blank (1991), finds a relatively large but imprecisely estimated effect of blinding on the acceptance rate of female-authored papers. Blank's findings are consistent with ours, but her sample lacked the power to precisely estimate gender differences. We return to the implications of our findings in the conclusions.

2 Data Extraction, Gender Coding, and Summary Statistics

2.1 Data Extraction

Our data are derived from information stored in the Editorial Express (EE) system used by each of the four journals. For confidentiality reasons we wrote a program that could be run by journal staff to create an anonymized data base, combining information in the EE system with gender information from pre-coded lists of author and referee names (see below). We are grateful to the four journals for agreeing to provide data access.

Our database builds on the submissions extract created by CDV in mid-2015. Google Scholar (GS) has created new barriers to accessing its data base in the past few years. We therefore elected to match our new data base back to the CDV data set and use the citations originally collected by CDV. We include all submissions from the first year each journal started using the EE system up to 2013 (as in CDV), leaving at least 1.5 years for citations to be realized.

Gender Coding. Prior to running our extraction program we obtained a list of the names of all authors and referees from each journal for the relevant years. We then developed a protocol for

assigning gender to the names on these lists. The protocol, laid out in Online Appendix Figure 1, relies on a combination of (1) public lists of given names including the fractions of people in the US and Germany with that name who are male; (2) lists of female economists' names; (3) a list of common Chinese given names; (4) internet search by a team of research assistants. We developed the protocol using a test data set of the names of 48,000 authors of articles published between 1990 and mid-2017 in a set of 63 economics journals (Online Appendix Table 1).

Our protocol begins by assigning "unknown gender" to common Chinese first names, since these names can be used by both males and females, and there are often multiple economists with the same Chinese name. This exclusion affects less than 1% of names. We then classify an author as **female** if *both* the US and German lists report that less than 1% of people with that first name are male, or if the full name is present in one of the lists of female economists. Likewise, we classify an author as **male** if one of the US or German names lists shows that over 99% of people with that name are male and the other shows at least 50% are male. Finally, a team of undergraduate research assistants looked up all names that remained unassigned.

Overall we are able to assign gender to about 97% of names in our test data set. As explained in Section 5.2, an audit of the genders assigned by our procedure for authors of 1,719 *published* articles in the four journals yields an error rate of under 1% (Online Appendix Table 13).

Analysis Plan. We posted an analysis plan on the AEA site under number AEARCTR-0003048 prior to the completion of our data collection. In our analysis, we follow the steps outlined in the plan, with the addition of a few robustness checks which we had not envisioned.

Survey. To help interpret our findings, we conducted a survey about perceptions of gender differences in the publication process. The survey was sent to three groups: (i) editors and co-editors at the 4 journals; (ii) a stratified random sample of 200 economists (100 male and 100 female) with at least 4 publications in our top-35 journal set from 2013 to 2017; (iii) all assistant professors of economics in the top 20 US schools and top 5 European schools with PhDs in 2015-17. The views of editors and co-editors are obviously relevant given their role in the publication process. The other two groups were selected to represent the views of economists with considerable recent publication experience, and of promising researchers at the start of their careers. Our response rates were reasonably high, ranging from 35% to 50% depending on the group.

Table I summarizes the responses to questions focusing on (i) how papers by mixed-gender teams are treated in the review process; (ii) whether female-authored papers are more likely to be assigned to female referees; (iii) how male and female referees evaluate male- and female-authored papers; (iv) the likelihood that the editor gives an R&R to male- versus female-authored papers; (v) the extent to which citations vary with author gender, holding constant the quality of a paper. Online Appendix Table 2 has additional information on the survey.

We use these answers in three ways. First, following our analysis plan, we use the beliefs about mixed-gender teams to motivate our classification into "senior female" and "senior male" teams. Second, we use beliefs about the potential differences in citations for male-authored and female-authored papers to help interpret the gaps in citations in our analysis. Third, we use the answers

as "priors" to benchmark our findings, as in DellaVigna and Pope (2018).

2.2 Summary Statistics

Table II presents summary statistics for the 15,147 submitted papers in our database that were not desk-rejected (NDR) and were assigned to at least two referees. We present summary statistics for the full set of 29,890 submissions in Online Appendix Table 3.[6]

We classify papers into five groups based on the author-gender mix: (i) all-male teams (67% of NDR papers); (ii) all-female teams (7%); (iii) mixed gender teams with a senior female co-author, i.e., the co-author with most prior publications is female (4%); (iv) other mixed gender teams (17%); and (v) gender undetermined, i.e., at least one co-author with unassigned gender (4%). Note that mixed gender teams in which none of the authors has any prior publications are assigned to the "other mixed gender" category, leading to a higher average publication record for the mixed gender teams with a senior female author than for any of the other groups (see below).

The top row of the table shows our benchmark measure of citations, the inverse hyperbolic sine ($asinh$) of GS citations, collected in mid 2015.[7] Citations are highest for mixed-gender papers (columns 3 and 4), followed by all-male papers (column 1), with all-female papers (column 2) at the bottom. We plot the cumulative distributions of $asinh(citations)$ by author-gender group in Online Appendix Figure 2a, and of residualized citations after adjusting for our key control variables (including prior author publications) in Online Appendix Figure 2b. The adjustment reverses the ranking, with higher residualized citations at most quantiles for female-authored papers.

Next we show the probability of an R&R verdict. Consistent with their relative citations, NDR papers by mixed-gender teams have the highest probability of an R&R, followed by all-male papers, with papers by female authors at the bottom. The editorial outcomes for all submissions (including desk rejections), summarized in Figure I Panel A, show a parallel pattern, with the lowest rates of desk rejection and highest unconditional rates of R&R for mixed-gender teams.

We measure the productivity of authors by the number of publications in 35 high-impact journals over the 5 years prior to submission (see Online Appendix Table 1 for the list of journals). On average across all authors in our submissions data, about 50% of males and 65% of females have no prior publications (see Figure I Panel C). We assign previous publications to each *paper* using the publication record of the most prolific co-author. As shown in Table II, authors of mixed gender papers have the highest numbers of prior publications. This is particularly true for mixed gender teams with a senior female co-author, though part of the gap is mechanical given how we assign papers for which none of the authors has any prior publications.

The number of coauthors is a key characteristic of papers that is highly correlated with author

[6] Among the non-desk-rejected papers, we exclude papers that were assigned to only one referee, since this process (which is especially common at the *Review of Economic Studies*) appears to be a form of desk-rejection.

[7] We use the $asinh$ transformation to accommodate zero citations. Bellmare and Wichman (Forthcoming) derive the elasticity of y with respect to x for a specification like $asinh(y) = \beta x + u$. For a continuous x this is $(\beta x/y)\sqrt{y^2+1} \approx \beta$x for $y \geq 2$. To address concerns about 0 citations, we present Poisson and Tobit models below, as well as models using $ln(1+y)$ as the dependent variable.

gender, since the likelihood of an all-female team of 2+ authors is relatively low in a field where only 20% of authors are female. Indeed, the rate of single authorship is over twice as high among all-female papers (72%) than all-male papers (36%) in the NDR sample.[8]

Another important characteristic of papers is field. We classify submissions into 13 broad groups based on their JEL code(s) and include separate codes for papers with no JEL codes, and for those that fall outside the main groups.[9] Author gender is correlated with field: female authors are more prevalent in the empirical micro fields like labor and development, and less prevalent in econometrics, finance, macro, and micro theory. Reassuringly, the share of female authors in different fields in our database is highly correlated with the share among papers in the EconLit database we use to validate our gender assignment process (Online Appendix Figure 3a), as is the share of female referees assigned to papers in a field (Online Appendix Figure 3b).

Since our data access agreement precluded the extraction of detailed JEL codes, we developed two measures of the gender-related content of different subfields. The first is the share of female authors of papers with each 2-digit JEL code published in the past 5 years in our EconLit sample of journals. The second is a simple indicator for JEL codes associated with gender-related topics, which we take to be D1 (Household Behavior and Family Economics), J1 (Demographic Economics), K36 (Family and Personal Law), and K38 (Human Rights Law and Gender Law). As Table II shows, all-female teams and mixed-gender teams with a senior female co-author are more likely to have JEL codes in "high female share" fields, and to include JEL codes indicating a gender-related topic.

Next we turn to the referee recommendations and characteristics of the referees. The summary recommendations in EE fall into 7 categories, ranging from "Definitely Reject" to "Accept".[10] Figure I Panel B shows the distributions of summary recommendations by author gender. Most papers receive a "Reject" or "Definitely Reject" recommendation. Again, mixed gender teams are more likely to receive a better review, while all-female author teams are less likely.

Finally, turning to the referees, on average about 15% are female. As is the case for authors, female referees tend to have fewer publications than male referees (Figure I Panel D).

3 Assignment of Referees

We begin our analysis by studying the matching process used by editors to assign NDR papers to referees. This tells us how editors treat different types of mixed gender teams, and provides some potential evidence on editors' concerns about gender-related issues in the review process.

As shown by the first set of vertical bars in Figure II, papers by all-female authors are assigned to female referees at nearly twice the rate (26 percent) as papers by all-male authors (14 percent).

[8] Interestingly, the fraction of single-authors who are female is very close to the fraction of authors of 2-author papers who are female, suggesting that females are no more likely to work alone. There is, however, some evidence of assortative matching of co-authors by gender. For example, among 2-authored papers the fraction written by two females is 4.1%, higher than the 2.5% rate expected under random matching.

[9] We classify a paper with J JEL codes as being assigned with weight $1/J$ to each of the listed fields. Some 11% of papers in our database do not have JEL information at submission and are thus coded as having missing field; most of these observations are due to two of the journals not collecting JEL information over 2 years.

[10] There are actually 8 categories, but "Conditionally Accept" and "Accept" are so rare we group them in one.

Mixed-gender author teams fall in between, with a higher rate for those with a senior female co-author (21 percent) than for the other mixed-gender papers (18 percent).

These simple comparisons do not account for other relevant paper characteristics, such as the field of the paper. We thus estimate linear probability models for the event that a paper is assigned to a female referee with our full set of controls, including journal-by-submission-year effects, the number of authors, the number of publications of the most prolific co-author, indicators for broad field, the share of female authors in the 2-digit JEL code(s) of a paper, and an indicator for gender-related subfields. We also control for the publication record of the referees assigned to each paper. We plot the residualized differences by author gender using the second set of bars in Figure II.[11] Compared to all-male authored papers, female-authored papers are 7 percentage points (s.e.=1 ppt) more likely to be assigned to a female referee, mixed-gender papers with a senior female co-author are 5 percentage points more likely, and other mixed-gender papers are only 3 percentage points more likely. The differences between the two mixed-gender groups motivate our choice to analyze these groups separately, along the lines described in our analysis plan.

One explanation for this gender-matching is that editors are concerned about reviewer biases. Indeed, the editors in our survey think that male reviewers are less likely to give a positive evaluation of female-authored papers than female reviewers (17.5% probability versus 20.7%). They are also more likely to expect gender-matching in the assignment of female authored papers than other groups in our survey. Another possible explanation (for which we have no evidence) is that female authors are more likely to cite other female authors – perhaps because they are working in an area dominated by female authors – and editors tend to select referees from a paper's bibliography. It is also however possible that the gender difference is due to subfield differences in the share of female authors and referees, which our field controls do not capture.

4 Referee Recommendations

4.1 Simple 'Audit' Comparison

The next question is whether referee gender actually affects the evaluations of papers. We use two main measures of referee support. The first summarizes the seven referee recommendations into an index based on the predicted $asinh(GS\,citations)$ associated with each category, using the coefficients from the main citation model in CDV (Table 2, Column 4). A second, simpler measure is the share of recommendations that are positive – that is, "Revise and Resubmit" or better.

Figure III Panels A and B compare the mean assessments by referee gender and author gender mix for papers assigned to at least one referee of each gender, with confidence intervals constructed by clustering at the paper level. We weight paper/referee observations by the inverse number of referees for a paper so that each paper receives equal weight. On average, female and male referees

[11] Online Appendix Table 4 reports the corresponding coefficients. We also examine, for comparison, the probability of being assigned to a referee with 3 or more recent publications. Controlling for other variables, female-authored papers are significantly less likely to be assigned to such referees, but the effect is small, a 5 percent reduction.

give very similar evaluations, tracking each other closely across author groups. There is no evidence of *a relative assessment gap* between male and female referees that depends on author gender.[12]

Table III presents a series of models at the referee-paper level, weighing by the inverse number of referees for the paper. With no additional controls (Columns 1 and 5), papers by all-female authors receive significantly lower evaluations, as do those by author teams with undetermined gender. When we add controls for field, number of authors, and previous publications (columns 2 and 6), however, we find no differences in how the referees assess all-female or mixed gender papers with a senior female author relative to all-male papers (the omitted group).

To test for any *differential assessment* of male- versus female-authored papers, we include paper fixed effects in columns 3 and 7, identifying the within-paper difference in assessments between male and female referees. By interacting referee gender with the author gender mix we identify the *differences-in-differences* in the evaluation of female versus male referees for papers by a particular author gender group, relative to all-male papers. In Columns 4 and 8 we add additional controls—interaction between referee gender and author publication, referee gender and field of the paper, and author gender mix interacted with referee publications—to ensure that the estimates are not biased by the fact that, say, female referees have fewer publications and referees with fewer publications are more likely to recommend an R&R for a particular author gender team.

Using any of these specifications, we find no large or statistically significant interactions between referee gender and author gender mix. Thus, there is no evidence of any *relative* bias by referees of one gender for or against papers by authors of the other gender.[13] This differs from the predictions of our survey respondents, who expected that female referees would be a bit more positive toward female-authored papers. In fact, if anything female referees (controlling for referee publications) are somewhat less likely to recommend positively on papers of all genders.

Our results are consistent, however, with findings in Abrevaya and Hamermesh (2012) who compare the effects of having a female referee for female-authored papers at an anonymous journal with double-blind refereeing, comparing submissions in 2000-08 (when internet searches could presumably reveal the author identities) to those in 1986-94 (when authors were more likely anonymous to the referee). Our estimated interaction effect between all-female authors and female referees in Column 8 (-0.007, s.e.= 0.031) compares to their "triple interaction" estimate of -0.01 (s.e.=0.08).

4.2 Model for Recommendations and Citations

Although the results in Table III rule out any large or statistically significant *relative bias* among male versus female referees, they do not necessarily imply that referees set the same standards for female and male authors. It is possible that referees of both genders are biased for, or against, female-authored papers. To make further progress we need to make comparisons *across* papers by different gender groups, taking into account differences in quality.

[12] We find a similar pattern if we use the full set of submissions up to 2017 (Online Appendix Figure 4a-b).

[13] Online Appendix Table 5 shows that these results do not appear to differ for more "senior" female referees (that is, referees with 3+ publications), versus the other female referees.

We proceed by using realized citations as an indicator of quality. Building on the simple model in CDV, we assume that each paper has true quality q which is only partially observed by editors and referees. At the R&R decision stage, the editor observes the gender composition of the author team, F, a set of other characteristics of the paper and author(s) x_1, and the referee recommendations R. For expositional purposes we assume that F and R are 1-dimensional, though in our analysis both are multi-dimensional. We assume that expected quality is a simple linear function of $x \equiv (x_1, F, R)$:

$$E[q|x] = \beta_0 + \beta_1 x_1 + \beta_F F + \beta_R R. \tag{1}$$

In addition, we assume that the prediction error $\phi_q \equiv q - E[q|x]$ is normally distributed with mean 0 and standard deviation σ_q.[14] Equation (1) allows for the possibility that the referees set higher or lower thresholds for assigning their recommendations based on the characteristics of the paper or its authors. If, for example, referees tend to give more negative evaluations to female-authored papers conditional on the factors included in x_1, then $\beta_F > 0$.

The second component of the model is the decision process of editors. We assume that the editor observes a noisy signal s of the error component ϕ_q, with $s = \phi_q + \zeta$, where $\zeta \sim N(0, \sigma_\zeta^2)$. Conditional on s and x, the editor's expectation of the paper's quality is:

$$E[q|s, x] = \beta_0 + \beta_1 x_1 + \beta_F F + \beta_R R + v, \tag{2}$$

where $v \equiv s \times \sigma_q^2/(\sigma_q^2 + \sigma_\zeta^2)$. With this forecast in hand, the editor then decides whether to reject the paper or invite a resubmission. For simplicity we assume that the editor makes a revise and resubmit decision ($RR = 1$) if $E[q|s, x]$ is above some threshold:

$$RR = 1 \Longleftrightarrow \beta_0 + \beta_1 x_1 + \beta_F F + \beta_R R + v > \tau_0 + \tau_1 x_1 + \tau_F F, \tag{3}$$

where τ_1 and τ_F incorporate editorial tastes for papers with different characteristics. In particular, if $\tau_F > 0$ the editor imposes a higher standard for female-authored papers.

Under our assumptions v is normally distributed, and the R&R decision is a probit model:

$$\begin{aligned} P\left[RR = 1|s, x\right] &= \Phi\left[\frac{\beta_0 - \tau_0 + (\beta_1 - \tau_1)x_1 + (\beta_F - \tau_F)F + \beta_R R}{\sigma_v}\right] \\ &= \Phi[\pi_0 + \pi_1 x_1 + \pi_F F + \pi_R R] \end{aligned} \tag{4}$$

where $\pi_0 = (\beta_0 - \tau_0)/\sigma_v$, $\pi_1 = (\beta_1 - \tau_1)/\sigma_v$, $\pi_F = (\beta_F - \tau_F)/\sigma_v$, and $\pi_R = \beta_R/\sigma_v$.

The third and final component of the model describes the determination of citations. We assume that cumulative citations c (measured some time after the determination of RR status) reflect a combination of quality and other factors summarized by η: $c = q + \eta$. One important

[14] A more complete model would start from an assumption about the formation of referee recommendations, conditional on (x_1, F), and derive the posterior distribution of quality conditional on (x_1, F, R). We note that if editors and referees have common priors and referees submit their posterior estimates of quality – so $R = E[q|x_1, F, read]$ – then we would expect $\beta_0 = \beta_1 = \beta_F = 0$ and $\beta_R = 1$.

factor included in η is the time from submission to the measurement of citations: we assume that this is incorporated in x_1. In addition, citations can depend on the gender of the author team, on the editor's R&R decision (which will determine whether the paper is published in the particular journal or not) and other random factors captured in an error component ϕ_η :

$$\eta = \eta_0 + \eta_1 x_1 + \eta_F x_F + \eta_{RR} RR + \phi_\eta. \tag{5}$$

Combining equations (5) and (1) leads to a simple model for citations:

$$c \;\; = \;\; \lambda_0 + \lambda_1 x_1 + \lambda_F F + \lambda_R R + \lambda_{RR} RR + \phi \tag{6}$$

where $\lambda_0 = \beta_0 + \eta_0$, $\lambda_1 = \beta_1 + \eta_1$,$\lambda_F = \beta_F + \eta_F$, $\lambda_R = \beta_R$, $\lambda_{RR} = \eta_{RR}$, and $\phi = \phi_q + \phi_\eta$.

A concern for estimation of the λ coefficients in equation (6) is that when the editor's signal is informative about future citations (so $\sigma_q > 0$ and $\sigma_\zeta < \infty$) RR status will be correlated with ϕ. In this case a simple OLS regression of citations on the variables (x_1, F, R) and RR status can lead to biases. To address this, we assume that different editors have different quality thresholds for reaching an R&R decision (i.e., different values of τ_0) but that the particular editor assigned to a paper has no direct effect on citations. This allows us to use the "leave out" R&R rate for the editor assigned to a paper as an instrumental variable for the editor's decision. Rather than estimating equation (6) by IV, we adopt a control function approach and include the generalized residual $\widehat{r}$ from a probit model for the editor's decision that includes x and the instrumental variable. Specifically, we estimate the augmented citation model:

$$c = \lambda_0 + \lambda_1 x_1 + \lambda_F F + \lambda_R R + \lambda_{RR} RR + \lambda_r \widehat{r} + \phi'. \tag{7}$$

As shown by Wooldridge (2015), the inclusion of $\widehat{r}$ absorbs the endogenous component of the R&R decision, yielding consistent estimates of λ_{RR} and the other coefficients of interest.

By comparing the effects of author gender on citations and the R&R decision, we can make an inference about the relative threshold that editors impose on female authors. As a starting point, suppose that $\eta_F = 0$, so the gap between citations and paper quality is unaffected by author gender. In this case the coefficients of R and F in the citation model are $\lambda_R = \beta_R$ and $\lambda_F = \beta_F$, while the corresponding coefficients in the R&R probit model are $\pi_F = (\beta_F - \tau_F)/\sigma_v$, and $\pi_R = \beta_R/\sigma_v$. If editors impose the same bar for male and female authors (i.e., $\tau_F = 0$) we would expect to find

$$\pi_F = \beta_F/\sigma_v = (\lambda_F/\lambda_R) \times \pi_R = \pi_F^{cite-max} \tag{8}$$

in the R&R model. In the absence of editorial preferences for or against female authors, and any gender bias in citations, the effects of the referee recommendations and the female indicator will be *proportional* in the models for citations and the editor's decision. The expected coefficient in the R&R probit under these assumptions is the same as the one that would emerge if editors selected papers based on the highest expected citations. We therefore refer to $\pi_F^{cite-max}$ defined in equation

(8) as the "citation-maximizing" benchmark for the coefficient of F in the R&R probit.

In contrast, if editors impose a *higher* bar on female authors (i.e., $\tau_F > 0$), then we would expect a systematic deviation from proportionality, with

$$\pi_F = (\beta_F - \tau_F)/\sigma_v < \pi_F^{cite-max},$$

i.e., a more negative effect of author gender on R&R rates. Conversely, if they impose a *lower* bar on female authors, then we would expect $\pi_F > \pi_F^{cite-max}$.

Suppose now that $\eta_F < 0$, so female authors get fewer citations than males, controlling for quality. In this case, the coefficient of F in the citation model is $\lambda_F = \beta_F + \eta_F < \beta_F$. If editors impose the same *quality* bar for male and female authors we would expect a coefficient for female authors in the R&R model that is *more positive* than under the citation-maximizing benchmark: $\pi_F = \beta_F/\sigma_v > \pi_F^{cite-max}$. Thus, if $\eta_F < 0$, the citation-maximizing benchmark is a *lower bound* on the expected coefficient for female-authored papers in the R&R model. In this case, a finding that the estimated coefficient in the R&R model is below the citation-maximizing benchmark can be interpreted as strong evidence that editors are not evaluating papers according to expected quality.

Gender Bias in Citations. To get a sense for the magnitude of any potential gender gap in citations, we asked our survey respondents to quantify their beliefs about the gap: "*Now consider two different papers in the same field of comparable quality, one written by female authors, the other written by male authors. Do you think the female-authored paper will get more, about the same, or fewer citations? If you answered more or fewer, how large do you think the citation difference will be in log points? For example, if you think that female-authored papers will have X log points (X percent) higher citations (conditional on quality), write X. If you think that female-authored papers will have X log points (X percent) fewer citations (conditional on quality), write -X.*"

While this question is admittedly a difficult one to answer, we found agreement on two points. First, the modal response across all groups is that there is no differential citation bias. Second, to the extent that there is a difference, respondents expect that female authors receive fewer citations than male authors (see Table I and Online Appendix Figure 5). The mean elicited citation bias is -6.5 log points: -3.9 log points among male respondents; -10.3 log points among female respondents; and -3.8 log points among editors.

4.3 Citations Conditional on Referee Evaluations

With this background, we consider the evidence on citations, starting with graphical evidence. In Figure IV Panel A, we plot average $asinh(citations)$ for each recommendation category, separately by author gender mix. We residualize citations with respect to all the controls, including the prior publications of authors.[15] At nearly each referee recommendation, female-authored papers have higher citations than male-authored papers, with a 20 log point average difference. This suggests

[15]Online Appendix Figure 6d shows the same graph with confidence interval. Online Appendix Figure 6a shows the evidence for the raw citation variable, with no controls.

that papers by all-female authors are held to a higher bar by the referees. This pattern is the same for male and female referees (Figure IV Panel B), consistent with the findings in Table III.

To aggregate across the recommendations of the different reviewers, we turn to the regression-based models in columns 1-4 of Table IV, following the format laid out in the analysis plan. We summarize the referee opinions by the fractions of recommendations in each of the 7 categories in Editorial Express. CDV document that this simple procedure provides a relatively accurate representation of the effect of the recommendations on both citations and the editor's R&R decision. To account for the facts that citations accumulate over time, and that average citations likely differ for papers submitted to different journals, we include journal $\times$ submission-year effects in all models.

In a specification that controls only for the referee recommendations and journal-year fixed effects (Column 1), female-authored papers receive 7 log points (s.e.=0.05) fewer citations than male-authored papers and mixed-gender papers receive 26-37 log points more citations. Once we add controls for prior publications, the number of co-authors, broad fields, and our two measures of gender-related fields (Column 2), however, the estimated citation premium for all-female teams rises to 24 log points (s.e.=0.06), while the premiums for mixed-gender teams fall and are insignificantly different from 0. These results show the importance of controlling for paper and author characteristics in making inferences about gender-related citation gaps.

We stress the interpretation of the results in Column 2. This specification controls for all observables, but *not* for referee recommendations. Thus, the 24 log point citation premium for all-female papers implies that, among all (non-desk-rejected) submissions, the papers with all-female authors appear to have higher quality, conditional on other observables.

How do these results change once we control for the referee recommendations in Column 3? All-female papers still have a significant citation premium of 22 log points, and mixed-gender papers continue to have small and statistically insignificant premiums relative to all-male papers.

As mentioned above, a confounding issue is that, to the extent that published papers get more citations, the effect of the referee reviews on citations could be biased. In column 4 we therefore add an indicator for a paper's R&R status and a control function to deal with the endogeneity of R&R status. The control function is the estimated generalized residual from the probit model for the R&R decision in column 7 of Table IV (and discussed below) that includes the leave out mean R&R rate of the editor who handles each NDR paper as an instrumental variable for R&R status.

This specification, which we take as our benchmark citation model, yields very similar estimates to the ones in the simpler model in column 3. In particular, the estimated citation premium for female-authored papers remains at 22 log points (s.e.=0.05). We interpret this gap as implying that a paper by an all-female team on average needs to have 25 percent ($exp(0.22) - 1 = 0.25$) higher citations (relative to a similar paper by an all-male team) to receive comparable recommendations. This gap is equivalent to the difference in citations between a paper that receives two "Weak R&R" recommendations and one that receives one "Weak R&R" and one "R&R" recommendation.

To the extent that female-authored papers tend to receive fewer citations than male-authored papers, the quality gap is even larger. Taking our estimate from the survey of a 6 log point gender

bias in citations, female-authored papers would have to be of 28 log points (32%) higher quality than male-authored papers to receive the same referee assessment.

Turning to mixed-gender teams, we find a relatively precise 0 citation gap for mixed-gender papers in which the senior author is male, consistent with the survey responses that indicate that such papers are considered similar to male-authored papers. For mixed-gender papers with a senior female author, the citation gap is 6 log point (s.e.=0.07). We cannot reject the hypothesis that these papers are treated "half-way" between female-authored and male-authored papers, the modal answer given by our survey respondents about how such papers are treated in the review process.

4.4 Robustness and Heterogeneity

Robustness. Three possible concerns with the estimated citation premiums for female-authored papers in our benchmark model are that (i) they may reflect the effect of some unobservable variable that is correlated with gender, (ii) they may depend on the functional form used for measuring citations, or (iii) they may be due to a peculiar subset of the data. We thus consider a broad spectrum of robustness checks in Table V and Online Appendix Table 6. For each specification (shown in a separate row), Columns 1-3 display the coefficients on the three author-gender variables from our citation regression, including the full set of controls as in Column 4 of Table IV. We discuss below the associated coefficients for the R&R decision, reported in Columns 4-6. For several of the robustness specifications, we report additional information in online appendix tables.

As far as unobserved factors, adding controls in Table IV significantly *increased* the estimated citation premium for all-female papers. If other unobservables tend to have the same correlations with female authorship and citations (as formalized in Altonji, Elder, and Taber, 2005), their omission would tend to downward-bias the estimate of the all-female citation premium. To provide more detail, Online Appendix Table 7 displays citation models with subsets of controls (including in all cases the referee recommendations and year-journal fixed effects). When we add only field controls, the results are similar to the specification with no controls. Adding the author publication variables shifts the coefficient on the all-female papers to 0.14 log points (s.e.=0.05), indicating that author publications are the key controls. Further adding controls for the number of authors raises the premium for all-female papers to 0.22, our benchmark specification.

Another possibility is that our controls for the publication record—the maximum number of publications in the 5 years prior to submission across the coauthors—are too crude. When we add controls for the *average* number of publications among the coauthors, for publications in top-5 journals (as opposed to 35 high-quality journals), and for publications 6-10 years prior to submission, the citation premium for all-female papers is unaffected. When we further add measures of the quality of the institution of the co-authors (reported also in row 1 of Table V), the citation premium for all-female authors falls to 0.17, while the premium for mixed-gender teams with a senior female coauthor falls to 0.04. On average, female authors are located at slightly more prestigious institutions, and papers by authors at higher-ranked institutions get more citations, so the addition of these controls lowers the estimated female author premiums by 2-4 log points.

A third possibility is that our field controls are not sufficiently detailed, and that females tend to write papers in particular sub-fields which tend to get more citations. Contrary to this line of reasoning, however, Table IV shows that both the average share of female authors in a sub-field and the share of sub-fields in gender-related areas have small, insignificant effects on citations. Still, to further address this, in the specification reported in row 2 of Table V (also reported in Online Appendix Table 7) we introduce interaction terms that allow the effect of the field variables to differ by year, and the effect of the gender-sub-field variables to differ by field. These additional controls slightly increase the female-author premium to 0.19 log points (s.e.=0.05). Taken as a whole, there is little evidence of an upward bias due to sub-field differences.

Next, in Online Appendix Table 8 we focus on alternative specifications of the citation model: (i) $ln(1+citations)$; (ii) a Poisson count model; (iii) the percentile of *citations* within the cohort of submissions to the same journal in the same year; (iv) an indicator for papers in the top x percent of citations, where x corresponds to the R&R rate in that journal-year cell (summarized in row 3 of Table V); (v) an indicator for "superstar" papers in the top 2 percent of citations for the journal-year cell; (vi) a Tobit specification for $asinh(citations)$, top-coding citations at 100. While one cannot directly compare the estimated citation premiums in these various models, we can compare their magnitudes to those of the referee recommendations. In our baseline model, the coefficient for an all-female authored paper is 12% as big as the coefficient for an R&R referee recommendation (0.22/1.89=0.12). In the probit model for being in the top x percent of citations, that ratio is similarly 12% and it rises to 22% in the probit model for being in the top 2% of citations, suggesting that female-authored papers have an even higher relative likelihood of achieving "superstar" status.

Two additional concerns about our citation measure are the role of papers with zero citations, and our use of Google Scholar rather than Social Science Citations Index (SSCI) citations. About 19% of all NDR papers in our data base have zero GS citations; it is possible that some of these papers may be recorded as zero citations due to, say, a special character in the author name or a typo in the title.[16] We deal with this in two ways: re-estimating the model excluding papers with 0 citations; and by modeling the left censoring of citations at zero using a Tobit model (columns 7 and 8 of Online Appendix Table 8). In both cases, the relative coefficient of all-female papers is 12% as big as the coefficient for an R&R referee recommendation, as in our benchmark specification.

Finally, we re-estimate the model using SSCI citations, which only include citations to published papers from other published papers. We limit the sample to submissions in the years 2006-2008 and fit a left-censored tobit model, given that, even among 2006-2008 submissions 60.8% have zero SSCI cites. This specification yields a 0.32 citation premium for all-female papers (s.e.=0.15), 16% as large as the estimated effect of an R&R recommendation.[17] Thus, our finding of a sizable positive citation premium for female-authored papers is robust to how we measure citations.

Heterogeneity. Table V and Online Appendix Table 6 also report a variety of estimates fit

[16] We investigate the possible correlation between author gender and missing citations using a subset of published papers in Section 5.2, and find no evidence of such a correlation.

[17] See column 10 of Online Appendix Table 8. For reference we also re-estimate our preferred model for $asinh(GS\ citations)$ using a Tobit model for the 2006-2008 period, yielding similar estimates.

to different subgroups. We estimate a larger female-author effect for papers from the earlier years (2003-2009, in row 6) than in the later years (2010-2013, in row 7). Citations to older papers are presumably less affected by conference presentations and prior circulation of working papers, so the larger female premium for older papers suggests that male authors may have short term advantages that if anything lead us to under-estimate the premium for all-female author teams.

In Online Appendix Table 6 (with the full results in Online Appendix Table 9), we estimate a citation premium for female authors of 17 log point with a single author and 34 points with 2 authors. For papers with 3+ authors we cannot reliably estimate the impact of an all-female team, given the rarity of such papers, but we estimate a 23 log point premium (s.e.=0.09) for mixed-gender papers with a senior female co-author.

In rows 9-10 of Table V we estimate our model separately for papers with fewer (0-3) versus more (4+) prior author publications. We find a positive citation premium for all-female author teams in both sub-samples, but the result is particularly large (49 log points, s.e.=0.10) for papers written by prolific teams. Unlike in Bohren et al. (forthcoming), we do not find a reversal of the pattern of gender differences for more highly experienced female authors.

Next, in rows 11-12 we estimate the model for sub-fields with above- or below-median shares of female economists. The citation gap for all-female papers is actually ***larger*** in fields with a higher share of female economists (like labor economics) than in fields with a lower share (like theory), though the difference is not statistically significant (t=1.4). This pattern is not supportive of the idea that the female-author premium arises from discrimination against female authors that is more prevalent in fields with fewer female authors.

Finally, we consider sub-samples based on referee characteristics. The gap is smaller for papers sent to only male referees (row 13) than for papers with at least one female referee (row 14) (though again the difference is not significant, t=1.3). Consistent with earlier findings, we see no evidence that male and female referees are differentially biased against female authors. As Online Appendix Table 6 shows, the female-authorship gap is larger for papers with 1 or 2 referees than for papers with 3 or more referees, though the difference is at best only marginally significant (t=1.7).

5 Editorial Decisions

So far, we have focused on how referees treat teams of authors with different gender compositions. But referees' opinions are only part of the editorial process: editors make the ultimate decision of whether to reject a paper or invite a revision. Moreover, editors make an initial screening decision on whether to desk-reject a paper or send it to referees. In this section we study the effects of author gender on editors decisions, again following our pre-analysis plan.

5.1 R&R decision

As we saw above, female-authored papers tend to get more citations at each level of the referee's recommendation (Figure IV Panel A). In contrast, the R&R rates are very similar for female- and

male-authored papers, conditional on the referee's opinion (Figure IV Panel C). This suggests that editors do not undo the relatively negative referee assessments of female-authored papers.

In Columns 5-7 of Table IV we fit a series of probit models for the R&R decision using the same variables as in our citation models. Our baseline specification in column 7 controls for the recommendations, the full set of controls, and the leave-out mean R&R rate of the editor assigned to the paper, included as an instrumental variable for the control function.

A comparison of the estimates in our baseline R&R model (column 7) and our baseline citation model (column 4) shows that the referee recommendation variables enter nearly proportionally, as would be expected if editors take the measures of referee support as an index of paper quality, and citations depend on the same index. Specifically, a plot of the R&R model coefficients for the 7 referee recommendation variables (the 6 reported in the table plus a 0 for the omitted category) against the citation model coefficients is approximately linear with a slope of 2.5. If editors are trying to maximize expected log citations, then all the variables in the R&R model should have coefficients that are 2.5 times larger than their coefficients in the model for citations.

Given that papers by all-female teams receive 0.22 log points more citations (Column 4), under a proportional decision model we would expect a coefficient of $0.55 = 0.22 \times 2.5$ in the R&R probit model, as specified by equation (8). A coefficient of this size would offset the bias in the referee recommendations and ensure that female-authored papers are evaluated in accord with their "quality", as revealed by citations. The actual coefficient on the all-female papers in the R&R decision in column 7, however, is 0.01 (s.e.= 0.06). Thus, as suggested by the graphical comparisons in Figure IV, editors do not *undo* the referee's apparent biases. The coefficient is precisely estimated, such that we can confidently reject the hypothesis of a value of 0.55 under the citation maximizing benchmark.

Note that equation (8) is derived under the assumption of equal citations for male and female papers of equal quality. Since most economists in our survey believe that female-authored papers receive if anything fewer citations for their work, equation (8) is actually a lower bound.

It is useful to draw a parallel to the case of the author publication variables, considered by CDV. Similar to the case of all-female authorship, having a well-published coauthor has a much smaller effect on the editor's R&R decision than would be expected under citation-maximizing behavior. For example, papers by authors with 6+ publications have an R&R probit weight of 0.41 (Column 7), compared to a predicted weight of $2.48 = 0.99 \times 2.5$. That is, author publications are underweighted by a factor of about 5; we cannot reject that all-female authorship is similarly underweighted by a factor of 5. A key difference, however, is in the interpretation. Authors with more publications could plausibly receive more citations for given quality of their work, reflecting more extensive networks and other factors. In the case of author gender, however, it seems implausible, at least to our survey responders, that female authors receive more citations than males for given quality.

Implications. How large is the impact of the non-proportionality of author gender in our citation and R&R models? We now simulate counterfactual R&R rates under two citation-maximizing counterfactuals.

In the first counterfactual, the editor corrects the deviation from citation maximization with respect to the author gender variables, but *not* with respect to the measures of author publications or other variables such as field. This is the relevant counterfactual if the editor believes that these other characteristics potentially affect citations, *conditional* on paper quality — for example, if citations to highly published authors are inflated. It is also the relevant counterfactual if, alternatively, there is no systematic bias in citations relative to quality, but editors set a higher bar for papers by authors with more publications as part of their editorial policy.

To compute the counterfactual, we start from the predicted R&R probability for the model in Column 7 of Table IV, $\Phi(x\widehat{\pi})$, which matches the R&R rate of 12.3% in the subsample of female-authored papers.[18] For all-female papers we then add the correction factor $\hat{\delta} = \widehat{\pi}_F^{cite-max} - \widehat{\pi}_F = 0.22 * 2.5 - 0.01 = 0.54$ and compute $\Phi(x\widehat{\pi} + F\hat{\delta})$. We add similar correction factors for the two sets of mixed gender papers. Since the resultant counterfactual R&R rates would lead to a change in the overall number of R&Rs, we adjust the journal×submission year effects in the R&R probit models to match the actual R&R rate (for NDR papers) in each journal-year cell.

As Figure V shows, the R&R rate for all-female papers would rise from 12.3% to 18.6% – a 50% increase. In contrast, under the counterfactual there is little effect for mixed-author papers. To compute the overall effect on the R&R rate for female economists, we average across the three groups of papers—all-female, senior female, and other mixed-gender papers— weighting by the number of female authors in each category. The R&R for an average female economist would increase from 14.2% to 15.9%, a 12% increase. This effect is large enough to potentially matter in career advancement and pay decisions.

In the second counterfactual, captured by the third set of bars in Figure V, we assume that the editor aims to maximize citations not just with respect to the author gender mix, but also with respect to the author publication variables. The R&R rate for all-female papers would remain roughly constant: these papers get a boost in R&R rates because of their gender, but they also suffer a relative reduction because female authors tend to have fewer prior publications, and the increased rate of R&R for highly published authors necessitates a raising of the overall bar for R&R status. Mixed-gender papers would benefit from the author publication correction, such that overall the R&R rate for papers by female economists would rise from 14.2% to 15.7%.

Robustness and Heterogeneity. In Online Appendix Tables 6-7 and 9-10 and in Table V we consider the robustness and heterogeneity of the R&R probit results, along the same dimensions considered for the citation models. Across virtually all specifications we replicate the key finding that editors fail to significantly upweight female-authored papers, despite the fact that female-authored papers get more citations. To illustrate this, in columns 7-9 of Table V we compare the actual R&R rate for papers by female economists in the particular sample or specification, and the counterfactual R&R rates according to the two citation-maximizing benchmarks. Across all specifications, the counterfactual would increase the R&R rate for papers by female authors, though

[18] In a slight abuse of notation we add the leave-out mean R&R rate of the editor assigned to the paper as a component of the vector x. We also note that a probit model does not necessarily predict the sample mean for a subgroup with an included dummy: in our case however, the deviation is negligible.

the increase is larger in some cases (e.g., papers submitted in years 2003-09) and smaller in others (e.g., papers submitted in years 2010-13).

5.2 An Analysis of Published Papers

So far, we have shown that all-female papers get more citations than all-male papers, conditional on the referee assessment, and that editors do not correct for this apparent bias in the referee assessments. There are two leading explanations for these results. The first is that referees and editors impose a higher bar on papers by female authors, perhaps because of stereotyping biases that lead them to downweight the quality of work by female authors. The second is that female-authored papers have characteristics that lead to higher citations but are not as highly rewarded in the review process. Examples could include working in particular sub-fields, or writing empirical, as opposed to theoretical, papers within a field. Given that we cannot adequately measure these characteristics in our main sample, it is difficult to separate out the two explanations.

To provide some evidence on the second explanation, we collected a secondary sample of all the published articles in the four journals of our sample for the years 2008-15. These 1,719 papers broadly correspond to the accepted papers in our sample, assuming a 2-year delay between submissions and publication. Although we have no information on the referee assessments of these papers, we are able to measure the gender and prior publications of each author (assuming the papers were submitted 2 years prior to publication). We also collected Google Scholar citations as of March 2019 for each papers, using the same procedures as for the main sample.

A key advantage of this sample is that we can measure a variety of features of each paper that are unavailable in our main sample, including detailed JEL codes. We are also able to study the measurement errors in citation counts arising from problems in our scraping procedure (associated with special characters in author's names and titles) and to estimate the misclassification rates of author gender. The key disadvantage is that it is about one tenth the size of our main sample, and it does not allow us to disentangle the role of referees and editors.

In Table VI, we first estimate our citation model (excluding referee evaluations) on the sample of *accepted* papers from our main EE database. We then repeat the same specification using published papers (column 2). The two samples are not identical (1,713 accepted papers versus 1,719 published paper) and have different citation measures (GS citations as of mid-2015 versus March 2019), but they have similar characteristics (Appendix Table 11). The estimated coefficients in the two samples are also similar. For the accepted papers in our main sample, the estimated female-author effect is 0.24 log points (s.e.=0.13), virtually identical to the estimate obtained in our overall sample when we exclude the referee variables (see column 2 of Table IV), though less precisely estimated. In the sample of published papers the female author citation premium is somewhat smaller at 12 log points (s.e.=0.11), but qualitatively similar. We also find a 24 log points (s.e.=0.13) citation premium for mixed-gender papers with a senior female coauthor in the published paper sample. These estimates replicate the key qualitative patterns in our main findings, and are qualitatively consistent with the findings of Hengel (2019) for a different sample of published papers.

We then add additional controls for paper features that are unavailable in the main sample. First, in Column 3 we include very detailed 2-digit JEL field controls (e.g., for J28), to assess the potential role of sub-field differences beyond our broad field controls. These controls nearly double the R^2 in the citation regression, but barely affect the gender coefficients. This concurs with the evidence above suggesting that sub-field differences are unlikely to explain the findings.

Next, we include two sets of measures of theoretical, as opposed to empirical, content. These are important additions because the JEL codes do not, by design, measure this dimension. To construct a first set of measures, we search the text of each published paper for words indicating theoretical content (such as "theorem" and "proposition"), empirical content (such as "standard error" and "table"), structural estimation (such as "structural" and "simulation"), or experimental component (such as "RCT" and "laboratory"). In Column 4, we add the asinh of the count of words in each of these four categories. The presence of theoretical material has a substantial negative predictive power on citations, and conversely for the presence of empirical results. Introducing these controls reduces the estimated author gender premia, reflecting the fact that papers by female authors and by mixed gender teams tend to have more empirical content.

The second set of measures are derived from a set of variables coded by a team of research assistants (who were not informed of our interest in gender-related differences). As shown in Column 5, these measures also show a negative citation effect for theoretical content and a positive effect for empirical content, and lead to somewhat smaller author gender premia. Adding both sets of variables (Column 6) leads to a citation premium of 9 log points (s.e.=0.13) for all-female authors, and 13 log points (s.e.=0.15) for mixed-gender papers with senior female authors.

In Column 7, we estimate a Lasso regression (using 10-fold cross validation), motivated by the fact that the specification on Column 6 has hundreds of sub-field controls and a dozen controls for other characteristics. The results are quite similar to the ones in Column 6, alleviating concerns about potential over-fitting of the latter model.

Papers by different author gender teams could also differ in the patterns of accrual of citations. While we cannot conduct such analysis with Google Scholar citations, in Online Appendix Table 12 we provide evidence using SSCI citation measures (which provide a full list of the citing articles). Specifically, we estimate models for the asinh of SSCI citations accumulated 0-1 years after publication, 2-3 years after publication, and 4-5 years after publication. We find no obvious pattern, with similar author-gender premia at 0-1 years out and at 4-5 years out, and smaller effects 2-3 years out. We also estimate models that separate out citations by the impact factor of the citing journals. We find slightly smaller gender premia using citations from high-impact factor journals, very small premia for citations at low-impact-factor journals, and relatively large premia associated with citations from journals with no known impact factor. While these results show no obvious pattern, we emphasize that the estimated gender premia are relatively imprecise.

Overall, the results on empirical versus theoretical content provide suggestive evidence that different characteristics of papers play some role in the observed citation premium, but would probably not explain all the citation premium in our main sample. Nevertheless, this conclusion is

tentative, given the difficulty of making inferences from just published papers.

Measurement Error. We hand-checked the citation counts and author gender for all papers in this sample. In terms of citations, we find that our automated scraper gave the correct citation count for 96% of papers (see Online Appendix Table 13). In 31 cases (1.8% of the sample) the title of the published article as recorded by Econlit had special characters, which led the scraper to find zero citations. Similarly, in 27 cases (1.6%) the author name(s) contained special characters (e.g., Jordi Galí), leading our automated scraping procedure to find 0 citations. The opposite type of error— finding too many citations—occurs in only 7 cases. Finally, with respect to gender coding, we found only 5 cases of incorrect coding of the author gender mix (0.3% of the sample), with a dozen additional cases in which we were able to assign gender to "undetermined-gender" cases.

The possibly consequential errors are the ones in which a paper is incorrectly recorded with 0 citations. If hypothetically, this error was less common for female economists, it could bias upward the measured citations of all-female papers in the main sample, potentially accounting for our key result. We see no reason *ex ante* why this error would correlate with author gender, given that it depends on special characters in the paper title or last name. Nonetheless, we provide two pieces of evidence. First, in Table V row 4 we run the main specification excluding papers with 0 citations and find similar results. Second, we measure whether the occurrence of papers with 0 citations differs by author gender, compared to the case of positive, but low, citations (defined as papers in the 35-55th percentile of citations in a journal-year). As Online Appendix Figure 7 shows, there is no obvious pattern by gender group and, to the extent that there is one, the all-female group has slightly *more* papers with zero citations, which would bias the citation measure downward.

5.3 Desk-Rejection

One interpretation of our results so far is that editors defer to the referees with respect to their R&R decisions. To provide evidence on editorial preferences with no input from referees, we turn to the desk-rejection decisions. Using the full sample of 29,890 submissions, we compare predictors of citations with predictors of the decision to not desk reject (NDR) in Table VII. At the submission stage, female-authored papers have 24 log points higher citations than submissions by all-male authors, holding constant other paper and author characteristics (Column 1). This mirrors the citation result in the R&R regression in Table IV.

As Column 2 shows, in the NDR decision editors *do* take into account the author gender and are more likely to not desk-reject papers by female authors, holding all else constant. This specification, though, does not tell us whether this is the optimal weight, given that we do not observe a variable like the referee recommendations.

For this, we build on a result in CDV: if the editors are putting the optimal weight on a variable X, that variable X should not predict citations once one controls for (a function of) the probability of desk-rejection. Thus, we re-estimate a citation specification including a cubic polynomial in $P(NDR)$ (from Column 2). We cannot include all the control variables, otherwise there will be essentially no identification left in the $P(NDR)$ cubic, but we do include at least the author

publication variable since CDV show that it is a strong predictor of citations, even controlling for the $P(NDR)$ polynomial. In Column 3 we include just this variable, while in Column 4 we also include journal-year fixed effects and controls for the female share in the sub-field of the paper. We estimate a smaller, but still sizable, all-female-author effect of 0.17 (s.e.=0.05) in Column 3 and 0.15 (s.e.=0.04) in Column 4, compared to 0.24 in Column 1. Thus, the desk-reject decision reduces the difference in citations by about a third, implying that the editors are only partially responding to the quality difference between female-authored papers and male-authored papers.

5.4 Weight Placed on Referee Recommendations

We now consider how editors use the information provided by male versus female referees. CDV find that the recommendations of more- and less-published referees are equally informative about the quality of papers (as measured by future citations). Yet, editors tend to place more weight on the recommendations of more published referees. Is there a similar difference by gender?

Figure VI Panel A shows the informativeness of referees (i.e. the relationship between referee recommendations and citations), by referee gender and prior publication record. (Referees with 3 or more recent publications are classified as "prominent"). The crucial element is the *slope* of these lines. Male and female referees do not seem to differ in their informativeness and, consistent with results from CDV, more and less prominent referees do not differ in their informativeness either. Figure VI Panel B shows the relative weight given by editors to recommendations from referees who differ in prominence and gender. Consistent with the results in CDV, more prominent referees are valued more, but gender does not seem to have much of an effect.

To estimate these patterns with controls, we consider a nonlinear model:

$$Outcome_i = \sum_{j=1}^{N_{referees,i}} (\alpha_0 Female_{ij} + (1 + \alpha_1 Female_{ij}) R_{ij}) / N_{referees,i}) + \boldsymbol{\gamma X_i} + \varepsilon_i$$

where $Outcome_i$ denotes paper i's outcome ($asinh(citations)$ or receiving an R&R decision), $Female_{ij}$ is an indicator for the gender of referee j of paper i, $N_{referees,i}$ is the number of referees who evaluate paper i, and R_{ij} denotes an index of recommendations $R_{ij} = \beta_{DefReject} DefReject_{ij} + \cdots + \beta_{Accept} Accept_{ij}$, with the same coefficients β_c for each recommendation type, regardless of the gender of the referee (or of the author team). Such a specification adjusts the index R_{ij} for both a gender-specific intercept and slope. If for example, female referees are more positive, we would expect a negative value for α_0 in the citation regression. If the recommendations of female referees are more informative, then we expect $\alpha_1 > 1$ in the citation regression.

Online Appendix Table 14 shows the results. Columns 1-3 have informativeness as measured by citations, while Columns 4-6 report models of the editor's R&R decision. Consistent with CDV, more prominent referees do not provide more informed recommendations (compared to less prominent referees), but nevertheless, the editors do value them more. Turning to gender, female and male referees do not differ in their informativeness, and editors do not put different value on

them either. This gender result is consistent with the expectations of the survey respondents.

6 Delays and Other Outcomes

We now turn to the speed of decision-making. If there are gender differences in the editorial process, they may appear in the form of decision delays.

In Columns 1-2 of Table VIII we consider referee-paper observations and estimate the number of days from paper submission to the report completion for reviewers who return a report. With controls for paper and author characteristics (Column 1), the response time of referees is not affected by the gender composition of the authors. In Column 2, we add paper fixed effects, fully controlling for possible differences in papers assigned to male or female referees, which may affect the time to report completion. We find that female referees are neither faster nor slower than male referees, a precise null effect. Further, there is no interaction with the gender composition of the authors.

To integrate the referee delays with the editor delays, in Columns 3-4 we consider decision times at the paper level, with the full set of controls as well as editor fixed effects. We detect no difference based on the author gender mix on the number of days from paper submission to the arrival of the last report (Column 3), or on the full decision time including editor delays (Column 4).

For the papers with an initial R&R that are ultimately accepted, we consider the impact of the author gender mix on the number of rounds of submissions (Column 5) and on the total time from submission to acceptance (Column 6), as well as the time that the authors take to submit the first revision and the number of days from the resubmission until the final acceptance (Online Appendix Table 15). Across all these specifications, we find no difference by the author gender mix.[19] Thus, unlike in the analysis of Hengel (2018), referee and editorial delays appear to be gender-neutral, a pattern also visible in Online Appendix Figures 8a-c.

We consider two further outcomes. In Online Appendix Table 16 we consider the propensity of referees to accept a referee invitation. Each observation is a referee request for papers that were not desk-rejected. Once we include controls (Column 2), we estimate no difference in the probability of accepting a referee invitation for all-female-authored papers. With paper fixed effects (Columns 3), we examine if female referee are more likely to provide public goods by agreeing to referee, but we find no such difference. Finally, we find no relative difference depending on whether the reviewer gender matches the author gender (Columns 3 and 4).

Finally, motivated by Hengel (2018)'s analysis, we examine in Online Appendix Table 17 the complexity level of the abstracts. An important caveat is that we only observe the abstract of the most recent version of the paper, and thus we cannot examine, as in Hengel (2018), the change in complexity from the submission to the published version of the paper; also, Hengel (2018) presents

[19] We find very similar results if we do not control for the referee recommendation variables. The finding that there is no difference in rounds of revision by author gender also provides suggestive evidence against an alternative interpretation of the citation premium for all-female papers—that male and female papers may be the same quality at the time they are read by referees, but female teams are more responsive to reports and thus their papers may eventually be better. This would presumably be reflected in shorter revisions.

a more in-depth analysis of abstract complexity. We simply compare the Gunning Fog (Columns 1 and 3) and Coleman-Liau (Columns 2 and 4) measures of complexity for papers with different author gender teams. We do so separately for the papers that were desk rejected or rejected (Columns 1-2), and the papers that received an R&R decision (Columns 3-4). We find no impact of the author gender mix on the readability of the abstract in either sample.

7 Discussion and Conclusion

Are the referees and editors in economics gender neutral? The answer is both "Yes" and "No".

If we focus on comparisons highlighted in previous research, we obtain relatively precise zero differences between gender groups. Considering referee recommendations, we replicate the findings of Abrevaya and Hamermesh (2012) that there are no difference in how referees of different genders assess papers by female and male authors. Considering delays in publication, as Hengel (2018) does, we find no differences by author gender. Turning to editorial decisions, which have not been previously studied, we find that editors are gender-blind in the sense that they treat female- and male-authored papers the same, conditional on the referee recommendations. Further, editors give about the same weight to recommendations of male and female referees, which is appropriate given that the two groups are equally informative.

Yet, the editorial process does not appear to be gender-neutral once we take into account underlying differences in paper quality, as revealed by future citations. Female-authored papers get 25 percent *more* citations than male-authored papers, controlling for other paper features including field and the authors' previous publication record. This estimate is relatively precisely estimated ($t > 4$) and is robust to a number of alternative specification choices. Given that any bias in citations as a measure of quality is likely to work against female authors, we interpret this finding as evidence that female researchers are held to a higher bar by referees (both male and female). Since editors do not adjust their thresholds for this higher bar, they effectively reject too many female-authored papers relative to a citation-maximizing benchmark.

What accounts for these patterns? We consider two main explanations. The first is that referees hold female authors to a higher bar, perhaps because of stereotype biases. The second is that female-authored papers have characteristics that lead to higher citations but are not as highly rewarded in the review process, such as being in particular sub-fields or having empirical content.

We turn to a set of published papers, broadly corresponding to the accepted papers in our sample, to further investigate this second explanation. The evidence suggests that sub-field differences are unlikely to play a role, but that differences in empirical content could play some role in the gender citation premium, though they would be unlikely to explain the full citation premium in our main sample. Nevertheless, this conclusion is tentative, given the difficulty of making inferences from just published papers. Future research with even more detailed information on paper characteristics for submitted papers will help further separate out these explanations.

Where does that leave us in terms of implications? Our finding that female authors appear to

be held to a higher standard is concerning. We estimate that as a result the R&R rate for papers with a female author is 1.7 percentage points lower than the rate consistent with a gender-neutral citation-maximizing rule. This gap suggests an important hurdle, aside from the assignment of credit in coauthored work stressed by Sarsons (2018), for junior female economists, as well as a continuing obstacle to career progression for more senior female researchers.

One potential remedy to help female economists – using more female referees – is unlikely to help, given that female referees hold female-authored papers to the same higher bar as male referees. Recruiting more female referees may only have the unintended consequence of requiring more public good provision by female economists (Babcock et al., 2017).

It appears to us that a simpler path is to increase the awareness of the higher bar for female-authored papers. The referees and editors can then take it into account in their recommendations and decisions. This would address the bias, whether its source is gender discrimination or undervalued paper characteristics. In contrast, a policy of double-blind evaluation, setting aside implementation difficulties, would address only the first source of bias.

It would be great to revisit our analysis in 3 to 5 years to test whether the gender difference has been corrected. Perhaps, as for NBA referee bias (Price and Wolfers, 2010), publicizing the findings may be enough to correct the pattern (Pope, Price and Wolfers, 2018).

This is just an example of the importance of data transparency in the editorial process, as CDV also stress. Indeed, we are grateful to the four journals in economics which agreed to such data access, something with very few parallels outside economics. The ability to systematically keep track of, and analyze, referee and editorial choices should make it relatively straightforward in the future to check for progress on any form of gender bias, especially if journals were to keep track of the gender of authors and referees in the editorial system. More generally, data transparency and access will help make the editorial process fairer and more efficient.

References

Abrevaya, Jason and Daniel S. Hamermesh. 2012. "Charity and Favoritism in the Field: Are Female Economists Nicer (To Each Other)?" *Review of Economics and Statistics*, 94(1): 202-207.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1): 151-184.

Azmat, Ghazala, and Rosa Ferrer. 2017. "Gender Gaps in Performance: Evidence from Young Lawyers." *Journal of Political Economy,* 125(5): 1306-1355.

Babcock, Linda, Maria Recalde, Lise Vesterlund, and Laurie Weingart. 2017. "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability." *American Economic Review*, 107(3): 714-747.

Bayer, Amanda, and Cecilia E. Rouse. 2016. "Diversity in the Economics Profession: A New Attack on an Old Problem." *Journal of Economic Perspectives*, 30(4): 221-42.

Bellemare, Marc F., and Casey J. Wichman. 2019. "Elasticities and the Inverse Hyperbolic Sine Transformation." *Oxford Bulletin of Economics and Statistics.*

Bertrand, Marianne, and Kevin F. Hallock. 2001. "The Gender Gap in Top Corporate Jobs." *Industrial and Labor Relations Review,* 55(1): 3-21.

Blank, Rebecca M.. 1991. "The Effects of Double-Blind Versus Single-Blind Reviewing: Experimental Evidence from the American Economic Review." *American Economic Review*, 81(5): 1041-1067.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. Forthcoming. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review.*

Bordalo, P., Katherine Coffman, and Nicola Genaioli. 2019. "Beliefs about Gender." *American Economic Review.* 109(3): 739-73.

Bransch, F., and M. Kvasnicka, 2017. "Male Gatekeepers Gender Bias in the Publishing Process?" IZA Discussion Paper No. 11089.

Broder, Ivy E. 1993. "Review of NSF Economics Proposals: Gender and Institutional Patterns." *American Economic Review*, 83(4): 964-970.

Card, David and Stefano DellaVigna. Forthcoming. "What do Editors Maximize? Evidence from Four Economics Journals." *Review of Economics and Statistics.*

Ceci, Stephen J., Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams. 2014. "Women in Academic Science: A Changing Landscape." *Psychological Science in the Public Interest*, 15(3): 75-141.

Chari, Anusha and Paul Goldsmith-Pinkham. 2017. "Gender Representation in Economics Across Topics and Time: Evidence from the NBER Summer Institute." NBER Working Paper No. 23953.

Christensen, Garret S., and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature,* 56(3): 920-980.

DellaVigna, Stefano and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies*, 85(2): 1029–1069.

Dolado, Juan J., Florentino Felgueroso, and Miguel Almunia. 2012. "Are Men and Women-Economists Evenly Distributed Across Research Fields? Some New Empirical Evidence." *SERIEs* 3: 367-393.

Donald, Stephen, G., and Daniel S. Hamermesh. 2006. "What is Discrimination? Gender in the *American Economic Association*, 1935-2004." *American Economic Review*, 96 (4): 1283-1292.

Ginther, Donna, K., and Shulamit Kahn. 2004. "Women in Economics: Moving Up or Falling Off the Academic Career Ladder?" *Journal of Economic Perspectives*, 18 (3): 193-214.

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104(4): 1091-1119.

Hengel, Erin. 2018. "Publishing while Female. Are Women Held to Higher Standards? Evidence from Peer Review." Working paper.

Hengel, Erin. 2019. "Gender Differences in Citations at Top Economics Journals," Working paper.

Hospido, Laura and Carlos Sanz. 2019. "Gender Gaps in the Evaluation of Research: Evidence from Submissions to Economics Conferences." IZA Discussion Paper No. 12494.

Lundberg, Shelly. 2017. "Report: Committee on the Status of Women in the Economics Profession (CSWEP)." *American Economic Review* 107(5): 759-76.

Niederle, Muriel, and Lise Vesterlund. 2011. "Gender and Competition." *Annual Review of Economics*, 3(1): 601-630.

Pope, Devin, Joseph Price, and Justin Wolfers. 2018. "Awareness Reduces Racial Bias." *Management Science*, 64(11): 4967-5460.

Price, Joseph and Justin Wolfers. 2010. "Racial Discrimination Among NBA Referees." *Quarterly Journal of Economics*, 125(4):1859-1887.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How Stereotypes Impair Women's Careers in Science." *Proceedings of the National Academy of Sciences*, 111(12): 4403-4408.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2015. "Taste for Competition and the Gender Gap Among Young Business Professionals." NBER Working Paper No. 21695.

Sarsons, Heather. 2018. "Gender Differences in Recognition for Group Work." Working Paper.

Wooldridge, Jeffrey M. 2015. "Control Function Methods in Applied Econometrics." *Journal of Human Resources*, 50(2): 420-445.

Wu, Alice. Forthcoming. "Gender Bias in Rumors Among Professionals: An Identity-based Interpretation." *Review of Economics and Statistics*.

A Online Appendix

A.1 Data

A.1.1 Data Extraction, Additional Details

Our data are derived from information stored in the Editorial Express (EE) system used by each of the four journals. For confidentiality reasons we wrote a program that could be run by journal staff to create an anonymized data base, combining information in the EE system with gender information from pre-coded lists of author and referee names (see below). The data agreement with the journals has two conditions: (i) no separate results by journal, and (ii) unlike the CDV data set, this supplemented data set will not be posted, even upon publication.

Our database builds on the submissions extract created by CDV in mid-2015. Google Scholar (GS) has created new barriers to accessing its data base in the past few years. We therefore elected to match our new data base back to the CDV data set and use the citations originally collected by CDV. Since CDV did not retain paper identifiers, we used a fuzzy match algorithm based on all the identifying variables stored in the CDV data base. This yields perfect matches for all non-desk rejected papers, but many-to-many matches for some desk-rejected papers. For papers with multiple matches, we calculate our primary measure of citations as a simple average of $asinh(citations)$ across all possible matches, though our results are virtually identical if we retain all possible matches and weight by the inverse number of matches for a given paper.

CDV only collected the publication record of the co-author with the most previous publications. We added information on publications of every co-author, as well as information on waiting times in the review process, on the gender composition of the sub-field of the paper, and on the complexity of the abstract.

A.1.2 Gender Coding, Additional Details

Our protocol for assigning names not included in other lists begins by assigning "unknown gender" to common Chinese first names, since these names can be used by both males and females, and there are often multiple economists with the same Chinese name. This exclusion affects less than 1% of names. We then classify an author as **female** if *both* the US and German lists report that less than 1% of people with that first name are male, or if the full name is present in one of the lists of female economists. Likewise, we classify an author as **male** if one of the US or German names lists shows that over 99% of people with that name are male and the other shows at least 50% are male. These cutoffs were derived using the Econlit test data set. Since only 20% of economists are female, we have to set higher cutoffs for assigning female gender than male to (roughly) equate the misclassification rates.

Finally, a team of undergraduate research assistants looked up all names that remained unassigned. If the assistant initially assigned to the name could not find a match, it was passed on to a second assistant. We had two separate assistants code a subsample of names that could not be assigned by the first-name procedure. The coders disagreed only 1% of the time; in 11% of cases neither coder was able to find a name match; and in 14% of cases only one of two coders found enough evidence to determine a gender.

A.1.3 Google Scholar Citations

We extracted Google Scholar (GS) citations using an automated web scraper. For every paper, we search the title of the paper in GS with "allintitle:" (e.g. "allintitle:Tagging and Targeting of

Energy Efficiency Subsidies"). This ensures that every result contains every word of the title stored in the Editorial Express (EE) archive. Then we extract the results of the first page of the search, retaining the list of authors and number of citations for each match. We then compare the last names of the authors reported for the GS match to the last names in EE, and keep search matches with at least one last name in the EE archive. Finally, we sum the citations of the matching entries to take account of earlier versions of papers that were circulated in different versions. Papers with no "allintitle" match in GS, or with no matching last names, are assigned 0 citations.

A.1.4 Survey, Additional Details

We conducted a survey of editors and economists, asking about their perception of gender differences in the publication process. The survey, approved under Berkeley IRB 2018-04-10955, was sent to three groups: (i) editors and co-editors at the 4 journals in our sample; (ii) a stratified random sample of 200 economists (100 male and 100 female) with at least 4 publications in our top-35 journal set from 2013 to 2017; (iii) all assistant professors of economics in the top 20 US schools and top 5 European schools with PhDs in 2015-17. Within each group, we did not keep track of individual respondents. Within the second and third group, however, we referred male and female respondents to different URL's to keep track of gender.

A.1.5 Published Papers

Sample. For the sample of published papers, we extracted papers in the *Journal of the European Economic Association*, the *Quarterly Journal of Economics*, the *Review of Economic Studies*, and the *Review of Economics and Statistics* from 2008 to 2015. To ensure a complete dataset, we extract the list of articles directly from the journal websites, for a total of 1838 articles. (We found that Econlit does not always provide the full list of papers published.) We obtain the JEL codes from Econlit. These JEL codes will in general differ from the ones in the main data set, which are entered at submission by the submitting author. The JEL codes recorded by Econlit overlap, but do not typically coincide with, the JEL codes on the journal website, as Econlit does its own assessment of field of the article.

From this data set, we exclude articles that fall under the following categories: Papers and Proceedings, comments, errata, corrigenda, notes, and editorial announcements and letters, yielding a final sample of 1719 papers. For each article, we download the PDF.

Using this sample, we run the same code as we used for our main sample to generate measures of author publications, author gender, number of authors, broad field, and the two measures of gender-field composition. We also use the same procedure to extract Google Scholar citations as of March of 2019. Summary statistics for this sample are reported in Online Appendix Table 11. We also extract the Web of Science (SSCI) citations, in this case downloading the data for all the citing papers (including publication year and journal), allowing us to perform the more detailed analysis of citations in Online Appendix Table 12.

Unlike in our main sample, we are able to check directly the accuracy of the key variables above. In particular, we hand check the Google Scholar citations and the gender coding of the authors for all papers in the sample. Online Appendix Table 13 presents the resulting evidence on the degree of measurement error in the data.

Additional Controls. For this sample of published papers, we build additional measures of paper characteristics which we do not have for the main sample. The first is a precise measure of sub-field of the paper. Specifically, for each paper, we obtain all the two-digit JEL code (i.e., G21) listed in EconLit. We create indicators for each individual JEL code, for a total of over 500

sub-field controls. We classify a paper with n JEL codes as being assigned with weight $1/n$ to each of the listed fields. For example, for a paper with 3 two-digit JELs, each of the three variables associated with those JELs are assigned $\frac{1}{3}$ and all other JELs are assigned 0.

Our second set of controls is designed to measure differences in the *content* of papers using counts of specific lists of words. We used an R program to search the full PDF version of each paper, assigning counts in 4 categories: theory, empirical, structural, and experimental. The list of words for theory content is: "proposition, theorem, lemma, proof, model, theory." The list for empirical content is: "empirical, data, standard error, table, regression, difference-in-differences, natural experiment, IV, RDD, impact, research design." The list for structural content is: "structural, logit, BLP, maximum likelihood, mixture, simulation, policy simulation, calibration"). Finally, the list indicating experimentally-based content is: "field experiment, RCT, laboratory, subjects, survey". We then use the inverse hyperbolic sin of the counts of words in each category as a measure of content in that domain.

The third set of controls also measures differences along the empirical/theoretical line, based on qualitative assessments by a team of undergraduate research assistants. Papers were randomly to 1 of 12 undergraduate assistants. The undergraduates were then asked to rank the roles of empirical analysis, modeling, and policy analysis in each paper on a scale of 1-10. Additionally, they were asked to count the number of propositions and theorems, the number of modeling equations, and the number of estimating equations. The specific instructions for the coding task were as follows:

For NoPropsTheorems, count the number of Propositions, Theorems, Lemmas, and Corollaries, including the Appendices if a part of the published paper. Do not count Claims or Definitions. If the Appendix includes a proof of a Proposition stated within the text, do not count this twice. Include Corollaries not in the text but present in any such proofs. For NoMathEquations, count the number of equations related to a "theoretical model," such as utility maximization, market equilibrium, etc. Count also equations that derive a theoretical econometric model. Include displayed equations but not those in footnotes. As in NoPropsTheorems, include equations in the published Appendix. Equations that span multiple lines count as one equation. For NoEstimatingEquations, count the number of equations related to empirical econometrics or estimation, such as OLS or IV specifications. If the equation is a derivation of a theoretical property of an econometric model, then include the equation in NoProsTheorems instead. For PagesofModel, count approximately the number of pages dedicated to modeling and deriving. Use integer numbers. For PagesofEmpirical, count approximately the number of pages dedicated to data or empirical model. Include any "Data Sections." Use integer numbers. For RoleofModel, provide a qualitative assessment of the role of model and theory in the paper on an integer scale from 0 (none) to 10 (absolutely central). Ask yourself, "How central is the modeling contribution in the paper?" For RoleofEmpirical, provide a qualitative assessment of the role of data and empirical methods in the paper, on an integer scale from 0 (none) to 10 (absolutely central). Ask yourself, "How central is the empirical contribution of the paper?" For RoleofPolicy, provide a qualitative assessment of the role and influence of policy in the paper, on an integer scale from 0 (none) to 10 (absolutely central). Ask yourself, "How relevant is this paper to policy?

As noted above, we collected information from Web of Science on the papers citing each of the published papers in our sample, including the journal in which it was published (some of which are outside economics). We then attempted to assign 5-year impact factors to the journals in which citing papers were published. In Online Appendix 12 Column 5 the subset of citing papers are those published in journals with an impact factor ≥ 5. In Column 6, the subset of citing papers are those published in journals with an impact factor < 3. Finally, the citing papers in in Column 7, are those published in journals for which we could not find an impact factor.

A.2 Counterfactual R&R Rates

Given the predicted R&R probability $\Phi(x_i\widehat{\pi})$ for paper i, we want to compute the counterfactual rate if editors were choosing papers to maximize citations. This requires finding what we call in the text $\pi_F^{cite-max}$, then finding a new set of journal×submission year constants for the R&R model such that the predicted probability of R&R across all submissions in the corresponding set of papers is equal to the actual probability.

To begin, we expand the model in section 4.2 by considering multi-dimensional vectors for the three gender groups with at least one female author: all-female authors, teams with a senior-female co-author, and other mixed gender author teams.[20] Specifically, define $\lambda_F = \left[\lambda_{fem}, \lambda_{sf}, \lambda_{mixed}\right]'$ and $\pi_F = \left[\pi_{fem}, \pi_{sf}, \pi_{mixed}\right]'$ as the vectors of coefficients in the models for citations and R&R, respectively, for gender teams with at least one female author.[21] As in the main text, define λ_R and π_R as the vectors of coefficients for the referees' evaluations in the two equations.

From equation 8, we know that in a citation maximization model, the coefficients on the gender variables in the citation model should be proportional to those in the R&R model, with a factor of proportionality equal to σ_ν. This is the same factor of proportionality as for the referee recommendation variables (see equation 4). We therefore estimate σ_ν by regressing the 7 coefficients of $\widehat{\pi}_R$ on the corresponding coefficients of $\widehat{\lambda}_R$. Call the resulting estimate $\widehat{\sigma}_\nu$. We then define a vector of *correction factors:*

$$\widehat{\delta}_F = \left[\widehat{\delta}_{fem}, \widehat{\delta}_{sf}, \widehat{\delta}_{mixed}\right]' = \pi_F^{cite-max} - \widehat{\pi}_F = \widehat{\lambda}_F \times \widehat{\sigma}_\nu - \widehat{\pi}_F$$

that represent the gaps between the citation-maximizing R&R coefficients for each gender group and their actual coefficients.

Finally, define $G_F = \left[g_{fem}, g_{sf}, g_{mixed}\right]'$ a vector of dummies that take value of 1 if a paper belongs to that gender group, and 0 if not.[22] We then compute the R&R rate under citation maximization with respect to gender variables as $\Phi\left(x\widehat{\pi} + G_F'\widehat{\delta}_F\right)$.

Note that the average predicted R&R rates implied by these corrected probabilities are too high, since R&R rates for papers with female authors are raised while those by papers with male authors remain the same. To overcome this issue, for every year-journal cohort c, we compute the correction factor ρ_c such that

$$\sum_{i \in c} \Phi(x\hat{\pi} + G_F'\widehat{\delta}_F + \rho_c) = r_c$$

where r_c is the average RR rate among the not desk rejected papers in a year-journal cohort.

In Table V, we repeat the above procedure for every specification we run, *i.e.* we compute the coefficient of proportionality $\widehat{\sigma}_\nu$ implied by each model, as well as the correction factors for fixed number of papers ρ_c. We note finally that this procedure can accommodate corrections with respect to any variable in the R&R and citation models. In Table V we show the results correcting for gender only and results correcting for both gender and the publication record of the author team.

[20] Note that we do not attempt to adjust R&R rates for the undetermined gender group. This group of papers, which receive few citations and have low R&R rates, is mainly comprised of papers with at least one coauthor that our research assistants could not find. We believe that their low R&R rates are explained by the fact that at least one coauthor has "disappeared" from professional research, rather than by gender of the coauthors

[21] The subsrict *fem* refers to all-female author teams, *sf* to mixed-gender teams with a senior female author, and *mixed* to mixed gender teams with a male senior author or no senior authors.

[22] Therefore both all-male authored papers and undetermined gender papers will have a vector of 0s.