# Gender and Refereeing in Environmental Economics

Prashant Bharadwaj, Yashna Nandan, Matthew Neidell, Sheila Olmstead *

January 29, 2024

**Abstract**

*Bharadwaj & Nandan: Department of Economics, UC San Diego; Neidell: Columbia Public Health; Olmstead: LBJ School of Public Affairs, UT Austin and Resources for the Future.

# I Introduction

Part of the gender gap in labor market outcomes, particularly in academic settings, can be attributed to women doing more unrewarded tasks (Babcock, Peyser, Vesterlund, and Weingart, 2022). In a recent article in the *Review of Environmental Economics and Policy*, Kuminoff, Ciaramello, Dooley, Heintzelman, Khanna, Kosnik, Lewis, and Trimble (2023) highlight the Association of Environmental and Resource Economists' (AERE's) commitment to diversity and introduce new data and benchmarks to document progress on diversity, equity and inclusion (DEI) issues within the field of environmental economics. Using data from 2000-2020, they focus on membership in AERE and add richness to conventional measures of DEI by looking at employer, alma mater, year of PhD, and other characteristics of those authors who published articles in the *Journal of the Association of Environmental and Resource Economists* (JAERE). An important point they make is that while women are well represented in proportion to economics PhDs in general in AERE (approximately 29%), they form a smaller share of authors in JAERE (approximately 16%), but a disproportionate share of AERE leadership (approximately 46%). Women may be sacrificing their private productivity by devoting more time to producing public goods, but at the same time these positions of leadership may render private benefits, including influence in the field and agenda-setting ability.

In this feature we add to the understanding of DEI within the field of environmental economics by examining gender differences in a ubiquitous task in academia that has a large public goods component and where private rewards are uncertain: refereeing. Refereeing for most journals in economics is an unpaid task, but it might be perceived as an important avenue for scholars to make a positive impression on editors who are often influential figures in the field, indicating potential private value to accepting requests and producing high quality reports. We use the entire universe of submissions to JAERE from 2014-2021 in our analysis, assigning gender based on names of editors and referees.[1]

Using manuscript fixed effects and controlling for referee experience, we find that male and female referees in JAERE are similar on several dimensions. An important difference is when referees are late submitting a report, female referees are late by fewer days, and this effect is more pronounced when the handling editor is more senior (measured as years since PhD). Since nearly all other dimensions of the report are the same (word count, sentiment, editor agreement with referee recommendation, etc.), we interpret this as female referees putting in more effort conditional on being late. Because perceived reputational

[1]We recognize that our binary definition of gender in this paper leaves out our non-binary colleagues in environmental economics. The number of referees and editors for whom we could determine non-binary gender identity was too small to allow for a separate non-binary category in the analysis. We leave this to future work as norms in the profession and in society evolve.

consequences are likely most salient along the dimension of lateness rather than other aspects of refereeing which are more subjective (e.g., rejection decisions, word count, etc.), this behavior is consistent with a model in which female referees perceive greater reputational costs for submitting late referee reports.

In addition to building on the excellent work of Kuminoff, Ciaramello, Dooley, Heintzelman, Khanna, Kosnik, Lewis, and Trimble (2023), we contribute more broadly to a recent literature examining gender differences within the economics profession. Several papers identify significant barriers for women in economics. For example, Ginther and Kahn (2021) show that women in economics are less likely to be promoted after controlling for several key observables, and Hengel (2022) and Hengel and Moon (2020) show that women are likely held to higher standards in publishing. Hence, women may rationally perceive higher reputational costs to submitting late referee reports. Within the realm of understanding gender dynamics in the refereeing process, Abrevaya and Hamermesh (2012) provide evidence that female referees are not more likely to provide positive reviews for papers with female authors. In our case, we focus on the dynamics between the referee and the editor, rather than the referee and the author. We find that female referees are less late for editors with more experience, even when controlling for the interaction between female referees and editor gender. This suggests that editor experience, rather than editor gender, differentially matters for female versus male referees.

# II Data

Our data include 1,538 paper submissions to JAERE between 2014 and 2021. JAERE has a team of four editors who share responsibility for assigning submissions, in two-month shifts. Manuscripts are received by the assigning editor on duty and are then either retained by that editor, or sent to another editor or co-editor to handle. The handling editor or co-editor then implements the process of inviting referees, reviewing reports, and using referee advice to make a final decision. Late notices and reminders are automated, written in standard wording, and sent by the journal's managing editor, a staff member at the journal's press. We define the handling co-editor or editor as the primary editor in all analyses. Note that the only compensation referees receive for their work is one complementary submission to the journal in the 12 months following submission of an on-time report (worth a $100 avoided submission fee).

The data contain detailed information covering the entirety of the refereeing process, from the date a referee is invited to review a manuscript to the date that the final decision is sent to the authors. Variables we observe include whether a referee is invited to review

a paper, the number of days a referee takes to respond to an invitation, the number of days a referee takes to complete a review, whether a referee submits their review before the deadline, a referee's recommendation, and the editor's final decision. A unique feature of our data is our ability to observe the actual content of referee reports and letters to the editor. When submitting their recommendations, referees have the option to submit their written reports either through a text-box entry along with their recommendation or as separate PDF attachments. Our current data include all text-box submissions as well the majority of PDF submissions for the years 2014-2021.

We observe full names of the authors, referees, and editors. We assign gender to referees and editors using the following algorithm: (1) An RA assigns perceived gender to all cases she is able to clearly identify using only the first name of the individual, (2) We then use the "genderize" program in R to assign gender to the sample of first names the RA was not able to identify, and (3) For the remaining names which cannot be identified solely by first name, an RA googles the full name and uses the profile of the economist to assign gender.[2] Using this method, we are able to assign gender to 1,458 out of 1,464 referees and to all 37 editors in the full dataset. We supplement these data with manually collected data on PhD completion years for both the referees and editors, obtained from their online CVs. This allows us to calculate the years since PhD completion for a referee or editor at any given year in the data. We obtain PhD completion years for 1,341 out of 1,464 referees and 35 out of 37 editors.

To study the content of reports, we measure the word count and sentiment of the referee reports and letters to the editor, respectively. We obtain our measure of sentiment using a rule-based method of sentiment analysis, which classifies each word in the cleaned text as positive or negative based on a dictionary of positive and negative words and returns a polarity score that ranges from -1 to 1, with 0 indicating neutral sentiment.[3]

Summary statistics by gender are presented in Table 1. In all analyses, we limit our data to non desk-rejected manuscripts in the initial submission round (i.e., round 0) and drop manuscripts in which more than one editor handled the manuscript between rounds. We also drop observations in which the referee gender is missing. This leaves us with a sample of 1,024 unique manuscripts, 1,441 unique referees, and 33 unique editors. Men make up a much larger percentage than women of both referees and editors. On average, referees have 14 years of experience, and editors have 19 years of experience, where years of experience is defined as the difference between the observation year and the year that a

[2]Since this approach may incorrectly assign gender identity, we interpret our measure as one of perceived gender, which may be equally relevant.

[3]Specifically, we use the "textblob" python library to conduct the sentiment analysis.

referee received their PhD.[4] Both male referees and male editors are more experienced than their female counterparts. On average, referees are invited to review a manuscript 1.4 times a year and 7 times overall during the sample years, with small differences between male and female referees. Focusing on the measures that will serve as our outcome variables, female referees are on average less likely to decline an invitation (although they receive a slightly lower number of invitations), less likely to reject, and less likely to submit their review late, though none of these means differ statistically from their male counterparts. On measures of referee report content, male and female referees exhibit even more similar behaviors. Any average differences we observe in the data may simply be explained by differences in the quality of manuscripts assigned to female and male referees - our econometric specification will utilize manuscript fixed effects to study *within* manuscript differences in referee behavior.

[4]For referees who are still PhD students at the time of invitation, we assign a value of 0 for years of experience.

Table 1: Summary Statistics

| | All | Female | Male |
|---|---|---|---|
| Manuscript Characteristics | | | |
| Number of Referees | 2.18 | - | - |
| | ( 0.58) | | |
| Share Female Referees | 0.20 | - | - |
| | ( 0.28) | | |
| Number of Authors | 1.68 | - | - |
| | ( 0.87) | | |
| Referee Characteristics | | | |
| Unique Count | 1441 | 339 | 1102 |
| Years Since PhD Completion | 13.78 | 11.64 | 14.35 |
| | ( 10.48) | ( 8.85) | ( 10.80) |
| Number of Times Invited Per Year | 1.42 | 1.37 | 1.43* |
| | ( 0.68) | ( 0.68) | ( 0.68) |
| Total Number of Times Invited | 8.01 | 6.87 | 8.31* |
| | ( 8.95) | ( 7.99) | ( 9.16) |
| Editor Characteristics | | | |
| Unique Count | 33 | 8 | 25 |
| Years Since PhD Completion | 18.91 | 15.50 | 19.98* |
| | ( 7.35) | ( 5.17) | ( 7.60) |
| Outcomes | | | |
| Reviewer Declined Invitation | 0.21 | 0.20 | 0.22 |
| | ( 0.41) | ( 0.40) | ( 0.41) |
| Reviewer Recommended Rejection | 0.57 | 0.53 | 0.57 |
| | ( 0.50) | ( 0.50) | ( 0.49) |
| Review was Late | 0.45 | 0.43 | 0.46 |
| | ( 0.50) | ( 0.50) | ( 0.50) |
| Number of Days Review was Late | 17.57 | 14.87 | 18.23* |
| | ( 16.36) | ( 13.39) | ( 16.94) |
| Editor Agreed with Referee | 0.75 | 0.76 | 0.75 |
| | ( 0.43) | ( 0.42) | ( 0.44) |
| Referee Report Wordcount | 1065.25 | 1121.29 | 1050.82* |
| | ( 588.80) | ( 637.81) | ( 574.83) |
| Referee Report Sentiment | 0.08 | 0.07 | 0.08 |
| | ( 0.05) | ( 0.05) | ( 0.05) |
| Letter to Editor Wordcount | 255.24 | 272.17 | 251.09 |
| | ( 289.73) | ( 305.60) | ( 285.65) |
| Letter to Editor Sentiment | 0.14 | 0.14 | 0.14 |
| | ( 0.13) | ( 0.11) | ( 0.13) |
| Unique Manuscripts | 1024 | - | - |

Notes: Sample includes non desk-rejected papers from 2014-2021 in the initial submission round (i.e., round 0). "Years of experience" is defined as the number of years since receiving a PhD. All outcome variables are summarized in terms of referee gender. An * is added when the t-test testing the difference in means is significant at the 5% level. Both "Referee Report Sentiment" and "Letter to Editor Sentiment" are measures that take values of [-1,1] and are calculated from a rule-based sentiment analysis of the respective texts.

# III Methods

Our primary specification to study the effect of referee gender on outcomes takes the form:

$$y_{rm} = \alpha + \beta FemaleReferee_{rm} + X_{rm} + \delta_m + \epsilon_{rm} \tag{1}$$

where $y_{rm}$ is the relevant outcome for referee $r$ in manuscript $m$, *FemaleReferee* is a dummy that equals 1 if a referee is assigned female, and $X_{rm}$ is a vector of controls at the referee-manuscript level, including referee years of experience, and year the referee was invited. We include manuscript fixed effects $\delta_m$ which allow us to control for differences in quality of papers assigned to male and female referees. Our specification thus allows us to test whether *within* a manuscript, referee behavior differs by gender.[5]

Although our specification addresses the concern of endogenous assignment of referees to manuscripts (e.g., if female referees are more likely to be assigned lower-quality manuscripts), it is possible that certain types of manuscripts are assigned to "harsh" reviewers (e.g., if female-authored papers are more likely to be assigned to referees with a higher propensity to reject). Without referee fixed effects, which we cannot include because there is no variation in *FemaleReferee* within a refereee, we cannot control for referee "harshness." This would be a concern if we were testing whether, for example, female-authored papers are more likely to be rejected. However, because our research question focuses on referee-editor interactions, rather than referee-author interactions, lack of referee fixed effects should not bias our results.

We then test whether any effects of referee gender are differential across editor characteristics. Specifically, we aim to understand whether female referees react to editor gender and editor experience differently than their male counterparts. To this end, we interact the referee gender dummy with a dummy for editor gender and a measure of editor experience in the following specification:

$$\begin{aligned} y_{rm} = \alpha + \beta FemaleReferee_{rm} + \gamma \left[FemaleReferee \times FemaleEditor\right]_{rm} \\ + \mu \left[FemaleReferee \times ExperiencedEditor\right]_{rm} + X_{rm} + \delta_m + \epsilon_{rm} \end{aligned} \tag{2}$$

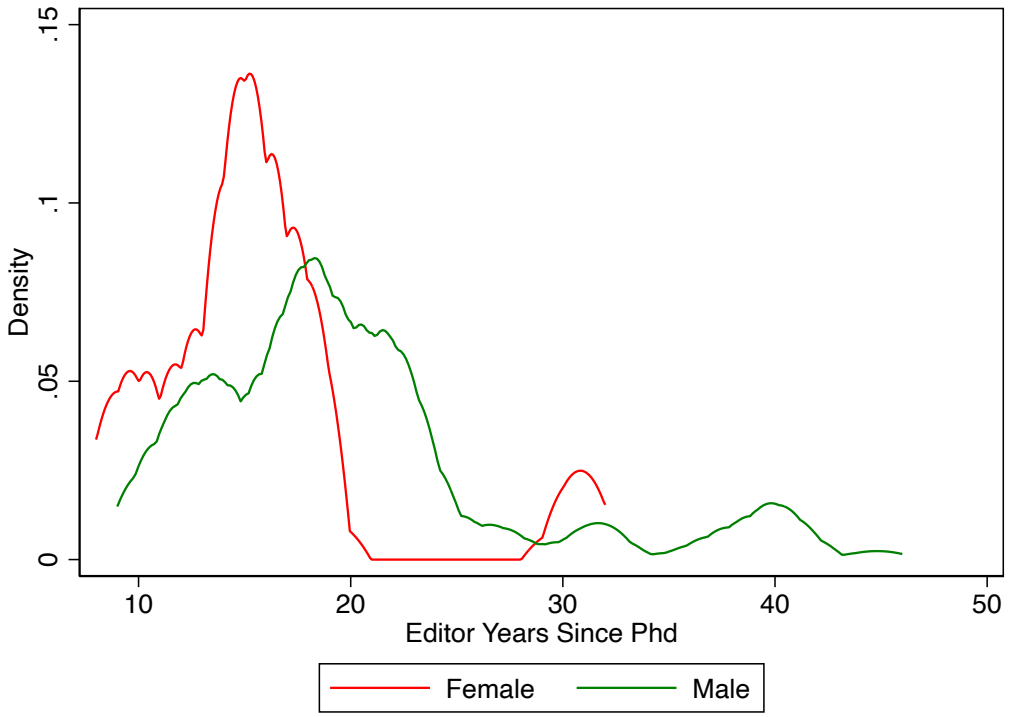
Here, *ExperiencedEditor* is a binary variable that equals 1 if an editor's years of experience is greater than the average value across the entire sample of editors and 0 otherwise. Figure 1 plots the distribution of editor experience for female and male editors. Female editors tend to be younger on average, while the most senior editors are all male. The correlation between editor gender and experience makes it important to understand whether interaction

[5]By using manuscript fixed effects, we drop all singleton observations. These are cases where only one referee is assigned to a manuscript or if, due to restrictions placed on the sample (e.g., conditioning on referees who were late), there is only one observation per manuscript.

effects between referee gender and editor gender are due to gender concordance itself or are actually driven by differences in experience. The coefficient $\gamma$ captures the differential effect of having a female editor for female referees relative to male referees. The coefficient $\mu$ thus captures the differential effect of having an experienced editor, for female referees relative to male referees, *separately* from any gender concordance effects.

Our primary outcomes are separated into four categories: declining behavior, refereeing style, report content, and editor agreement. For declining behavior, we use a binary indicator that equals 1 if a referee declined an invitation and 0 otherwise. Our measures of refereeing style include whether or not a referee recommended rejection, a binary indicator for whether the report was late, and the number of days that the report was late (conditional on being late). We explore additional measures of declining behavior and refereeing style in Appendix Table A1 and Appendix Table A2. Specifications that examine differences in the content of referee reports use the sentiment analysis variable defined earlier, ranging from -1 to 1, with 0 indicating neutral sentiment.

Figure 1: Distribution of Editor Experience by Gender




Notes: Figure plots distribution of editor experience, defined as years since receiving a PhD.

# IV Main Results

## Declining Behavior

We begin by analyzing the propensity to decline an invitation by referee gender in columns 1 and 2 of Table 2. We include controls for the year a referee was invited and the number of years since PhD completion, and cluster standard errors by manuscript. Column (1) presents our results from estimating equation (1). We find that female referees are no less likely to decline an invitation to referee. Additionally, the coefficient of 1.1 percentage points is small relative to the mean of 21.5 percent, and it is positive, indicating females are, if anything, more likely to decline. In column (2), we add interactions with editor gender and editor experience. We find that female referees are not differentially more or less likely to decline an invitation when their editor is female or experienced, relative to male referees. In Appendix Table A1 and Appendix Table A2, we explore additional outcomes relating to the initial invitation to referee, such as the number of days taken to agree, number of days taken to decline, and the number of invitations sent, and find generally similar patterns. Note that this null result addresses any concerns of a sample selection problem in subsequent analyses. If, on the other hand, we had found that female referees are less likely to decline the initial invitation, we may have been concerned that any differences we find in referee behavior were driven by differential selection into the sample.

## Refereeing Styles

In columns 3-8 of Table 2, we test whether female referees behave differently *within* the refereeing process. We include controls for the year a referee was invited and the number of years since PhD completion, and standard errors are clustered by manuscript. In column (3) we find no significant difference in rejecting behavior between female and male referees. The point estimate of 3 percentage points is small relative to the mean of 55.2. Additionally, female referees are not differentially more or less likely to reject when the editor is female or experienced. Moving to the time taken to submit referee reports, on average, female and male referees do not differ in *whether* they submit a report late (columns 5). However, as shown in column (6), when the editor is highly experienced, female referees are 14.4 percentage points less likely to be late relative to male referees, though the estimate is marginally significant. In columns (7) and (8), we limit the sample to referees who submitted a late report. Conditional on being late, we see that female referees take roughly 8 fewer days to submit their report relative to male referees. Relative to the mean of 17 days, this is a large and meaningful effect. Adding interactions with editor characteristics in column (8), we find that conditional on being late, relative to male referees, female referees take 15 fewer days

to submit their report when the editor is highly experienced. However, female referees do not take differentially less time to submit their report when the editor is female, suggesting that gender concordance is not driving gender differences in lateness.

Given our ability to observe the text contained in referee reports and letters to editor, we are able to study whether the content of these reports differ based on referee gender. Note that there are cases in which referee reports and letters to the editor are missing in our data. Thus, our sample size drops to 1,750 observations for referee reports and 1,346 observations for editor letters. We show these results in Table A3 and Appendix Table A4. The broad takeaway is that along a host of content-related measures such as word count, sentiment, female and male referees are quite similar and there is no difference by editor characteristics.

## Editor Agreement

To understand whether referee gender differentially influences editor decisions, we test whether an editor follows a referee's recommendation in columns 9 and 10 of Table 2. We include controls for the year the referee was invited and referee years of experience. Standard errors are clustered by manuscript. Note that 75% of the time, editors follow a referee's recommendation, but editors are no more or less likely to follow the recommendation of a female referee than that of a male referee. The point estimate of 3.0 pp is also small relative to the mean of 75 %. Again, we find no large or significant interaction effects of editor gender or editor experience with referee gender. This suggests that although female referees complete their reports more quickly and potentially provide more negative content, editors do not consider their recommendations more informative than those of their male counterparts. Similarly, although conditional on being late, female referees submit reports faster than male referees when the editor is experienced, experienced editors do not consider female referee recommendations to be more informative than male referee recommendations.

Table 2

| | Reviewer Declined | | Reject | | Review was Late | | # Days Late | | Editor Agreed with Referee | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| referee is female | 0.011 | 0.029 | -0.030 | -0.069 | -0.035 | 0.005 | -7.596*** | -3.998 | 0.030 | 0.020 |
| | (0.022) | (0.033) | (0.035) | (0.053) | (0.036) | (0.052) | (2.552) | (3.649) | (0.035) | (0.053) |
| editor and ref are female | | -0.037 | | 0.037 | | 0.017 | | 2.119 | | -0.011 |
| | | (0.052) | | (0.079) | | (0.083) | | (4.938) | | (0.080) |
| ref is female x editor is experienced | | -0.026 | | 0.090 | | -0.144* | | -15.212** | | 0.041 |
| | | (0.050) | | (0.080) | | (0.083) | | (6.516) | | (0.080) |
| Observations | 3111 | 3111 | 1942 | 1942 | 1942 | 1942 | 544 | 544 | 1942 | 1942 |
| $R^2$ | 0.298 | 0.298 | 0.527 | 0.527 | 0.540 | 0.541 | 0.559 | 0.578 | 0.375 | 0.376 |
| controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Manu FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean | .215 | .215 | .552 | .552 | .459 | .459 | 17.228 | 17.228 | .752 | .752 |

Notes: Observations are unique at the referee-manuscript level. The sample includes all non desk-rejected papers. "# of days late" is conditional on a review being late (i.e., late=1). "Editor Agreed with Ref" takes a value of 1 if the editor followed the referee's recommendation in a given round and 0 otherwise. "Editor is experienced" is a dummy variable that takes a value of 1 if an editor's years of experience is greater than the average value among all editors in the sample, and 0 otherwise. Regressions control for year invited and referee experience. Standard errors are clustered by manuscript.
* p<0.10, ** p<0.05, *** p<0.01.

## Multiple Hypothesis Testing

We study a total of nine outcomes in our main analysis. To address concerns of multiple hypothesis testing, we use two methods of multiple hypothesis adjustments : Anderson's False Discovery rate (FDR) sharpened q-value method and the Westfall-Young (WY) method. The FDR method limits the *proportion* of type 1 errors while the WY method limits the probability of making *any* type 1 error. In Table 3, we present original and adjusted p-values for all primary hypotheses tested in specification (1) and (2). Adjusted p-values account for the number of hypotheses tested in each specification. For example, in column (3) of Table 3, we use the Westfall-Young procedure to adjust the p-value for the coefficient on "Ref is fem" to account for the nine main outcomes in specification (1). In column (4) of Table 3, we use the Westfall-Young procedure to adjust the p-value for the coefficient on "Ref is fem X Ed is experienced" to account for the nine outcomes tested in specification (2). We follow the same procedure in columns (5) and (6) but use Anderson's FDR method to construct sharpened p-values.

We find that adjusting for MHT results in a loss of significance for our interaction term between referee gender and editor experience. However, the significant main effect of referee gender is intact. We also find that small differences by gender in word count and sentiment are no longer significant after adjusting for MHT.

Table 3: Multiple Hypothesis Testing

| Outcome | Original P-Value | | Westfall-Young P-Value | | Sharepened Q-Value | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | Ref is fem | Ref is fem X Ed is experienced | Ref is fem | Ref is fem X Ed is experienced | Ref is fem | Ref is fem X Ed is experienced |
| Reviewer Declining | | | | | | |
| Declined | 0.617 | 0.606 | 0.941 | 0.976 | 1.000 | 1.000 |
| Refereeing Styles | | | | | | |
| Reject | 0.379 | 0.257 | 0.905 | 0.874 | 0.644 | 0.947 |
| Review was Late | 0.330 | 0.084 | 0.905 | 0.512 | 0.644 | 0.507 |
| # Days Late | 0.003 | 0.020 | 0.033 | 0.185 | 0.030 | 0.224 |
| Report Content | | | | | | |
| Ref Report wordcount | 0.076 | 0.994 | 0.429 | 1.000 | 0.230 | 1.000 |
| Ref Report Sentiment | 0.035 | 0.332 | 0.257 | 0.897 | 1.000 | 0.947 |
| Letter Wordcount | 0.978 | 0.315 | 1.000 | 0.897 | 0.178 | 0.947 |
| Letter Sentiment | 0.981 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 |
| Editor Agreement | | | | | | |
| Agree | 0.391 | 0.608 | 0.905 | 0.976 | 0.644 | 1.000 |
| Specification | 1 | 2 | 1 | 2 | 1 | 2 |

Notes: Table presents original and adjusted p-values for the primary hypothesis tested in equations (1) and (2). Adjusted p-values account for the number of primary hypotheses tested in each specification. We test nine outcomes in each specification, so we adjust for nine hypothesis tests for each specification. Odd numbered columns present the p-values associated with "Referee is Female" coefficient in equation 1. Even-numbered columns present the p-values associated with " Referee is Female X Editor is Experienced" coefficient in equation 2.

# Discussion and Conclusion

In this feature, we expand on recent work by Kuminoff, Ciaramello, Dooley, Heintzelman, Khanna, Kosnik, Lewis, and Trimble (2023) to study gender differences in a traditionally unpaid task with large public returns: academic refereeing. Using data on the refereeing process for JAERE, a top environmental economics field journal, we employ a research strategy that uses manuscript fixed effects to look at differences in referee behavior within a manuscript and how this differs based on editor characteristics. We find that women are no more likely to agree to referee a paper for JAERE when asked than are their male counterparts, thus our results are not consistent with women agreeing to take on more unrewarded tasks, at least in this small setting. Along several other dimensions (e.g., word count, rejection, editor agreement with the referee's recommendation), male and female referees exhibit similar behavior.

However, conditional on submitting a late report, female referees submit their reports much faster than their male counterparts, and this effect is particularly strong when the editor is highly experienced. Specifically, we find that conditional on being late, female referees submit their reports in 8 fewer days than male referees, a large effect relative to the mean of 17 days late. When the editor is experienced, female referees submit their late report in 15 fewer days than their male counterparts. Importantly, we find that these interaction effects between referee gender and editor experience hold even when controlling for gender concordance effects. We interpret these findings as being consistent with a model where female referees perceive a greater reputation cost to lateness than male referees, and where this cost is greater when the editor is highly experienced. Why would reputational concerns only affect lateness and not other dimensions of refereeing? Other measures of refereeing, such as rejection recommendation, word count, and sentiment are inherently subjective and difficult for an editor to evaluate. However, lateness of a report is clearly negative and, in fact, well documented within journal reviewing systems, such as that used by JAERE. Thus, it is reasonable to assume that referees would perceive greater reputational costs to lateness and exert effort along this dimension. Future work that directly measures differences in perceived reputational costs by gender and tests whether female referees face higher reputational costs to lateness would make valuable contributions to our understanding of the dynamics of refereeing in economics.

# V Appendix: Robustness and Additional Outcomes

## Additional Outcomes

In this section of the Appendix, we explore additional measures of declining behavior and refereeing styles not presented in our main analyses. In Table A1, we study the intensive margin of declining behavior, focusing on the number of days to accept a referee invitation conditional on agreeing, the number of days taken to decline conditional on declining, and the number of invitation reminders sent by the editor (unconditional on accepting or declining). Note that the number of observations for the first two outcomes does not sum to the number of observations for the third outcome because the dates of agreement and decline are missing for several observations. In columns (1) and (2) we test whether female referees take longer to accept an invitation, conditional on agreeing. We find that female referees take 0.76 more days to accept an invitation, which is large relative to the mean of 2.7 days. However, female referees do not take differentially more or less time to accept when the editor is female or highly experienced (column 2). Conditional on declining an invitation, female and male referees do not differ in the number of days they take to decline, and we find no evidence of interactions between referee gender and editor characteristics in this measure (columns (3) and (4)). Turning to the number of invitation reminders sent by the editor, female referees receive 0.068 more invitation reminders, which is a 35% increase relative to the mean value of 0.194. This is consistent with the fact that female referees take significantly longer to accept an invitation. However, when matched with a highly experienced editor, female referees receive 0.096 fewer invitation reminders compared to male referees.

We next explore additional outcomes relating to refereeing styles. In our primary analysis of refereeing style in Table 2, we focus on extensive and intensive margins of lateness, but in columns (1) and (2) of Appendix Table A2, we study the total number of days to submit a review, conditional on on-time report submission. In column (1), we find no evidence that female and male referees differ in the total days taken to submit a review. However, once we add interactions between referee gender and editor characteristics in column (2), we find suggestive evidence that when the editor is female, female referees take 6.1 days longer than male referees. In columns (3) and (4), we again limit the sample to referees who are late, and test how the number of reminders sent by the editor differs by referee gender. Consistent with the evidence that conditional on being late, female referees take fewer days to submit their review, we find that female referees are sent 0.97 fewer reminders after the due date, a large effect relative to the mean of 1.87 reminders. Additionally, when the editor is highly experienced, female referees are sent 1.5 fewer reminders than their male counterparts, again consistent with the large interaction effect we found between referee gender and editor experience in Table 2.

Table A1: Declining - Additional Outcomes

| | Days to Agree | | Days to Decline | | # Invitation Reminders Sent | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| referee is female | 0.763** | 1.043 | -0.345 | 0.576 | 0.068*** | 0.104*** |
| | (0.382) | (0.734) | (0.758) | (1.189) | (0.023) | (0.032) |
| editor and ref are female | | -0.743 | | -0.736 | | -0.025 |
| | | (0.862) | | (1.665) | | (0.052) |
| ref is female x editor is experienced | | -0.223 | | -2.440 | | -0.096* |
| | | (0.876) | | (1.770) | | (0.051) |
| Observations | 2099 | 2099 | 383 | 383 | 3111 | 3111 |
| $R^2$ | 0.477 | 0.478 | 0.587 | 0.590 | 0.329 | 0.330 |
| controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Manu FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean | 2.724 | 2.724 | 3.172 | 3.172 | .194 | .194 |

Notes: Observations are unique at the referee-manuscript level. The sample includes all non desk-rejected papers in the initial submissions (i.e., round 0). "Days to Agree" is the number of days a referee takes to accept an invitation conditional on accepting, "days to decline" is the number of days a referee takes to decline an invitation conditional on declining. "# of Invitation Reminders Sent" is the number of reminders an editor sends between invitation and the referee's decision to accept or decline. "Editor is experienced" is a dummy variable that takes a value of 1 if an editor's years of experience is greater than the average value among all editors in the sample, and 0 otherwise. Regressions control for year invited and referee experience. Standard errors are clustered by manuscript. * p<0.10, ** p<0.05, *** p<0.01.

## Robustness of Sentiment Measure

In our analyses of referee report content, we present measures of sentiment calculated using a commonly-used method in rule-based sentiment analyses. In this section, we repeat our analysis using an alternative measure that uses a different dictionary to classify words as positive or negative and outputs a score with a range of [-1,1]. This measure is optimized for social media text and was thus not chosen as our primary measure. In Table A4, we find that on average, using this alternative measure produces much more positive sentiment scores for both referee reports and letters to the editor (0.82 and 0.75 compared to 0.08 and .14 using the original measure). Although this is a very large difference in average sentiment across measures, given that the alternative measure is tailored to social media text, it is plausible that words used in referee reports would be more likely to be classified as "positive" using this dictionary. However, even using this alternate score, we find that once we condition for interactions between referee gender and editor characteristics, referee reports written by female referees are slightly more "negative", although the effect size of 0.09 is small relative to the mean of 0.79. In contrast to our original findings, we see that letters to the editor written by female referees are also slightly more "negative", again with a small effect size. Similar to our original findings, we find no evidence that female referees are differentially more negative in their writing when the editor is female or highly experienced.

## Table A2: Refereeing Styles - Additional Outcomes

| | # of Days to Submit Review | | # Reminders Sent After Due Date | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| referee is female | 0.914 | -1.104 | -0.965*** | -0.779* |
| | (1.533) | (2.049) | (0.319) | (0.467) |
| editor and ref are female | | 6.087* | | 0.836 |
| | | (3.634) | | (0.582) |
| ref is female x editor is experienced | | 1.228 | | -1.546* |
| | | (3.672) | | (0.799) |
| Observations | 710 | 710 | 544 | 544 |
| $R^2$ | 0.628 | 0.630 | 0.559 | 0.578 |
| controls | Yes | Yes | Yes | Yes |
| Manu FE | Yes | Yes | Yes | Yes |
| Mean | 29.569 | 29.569 | 1.869 | 1.869 |

Notes: Observations are unique at the referee-manuscript level. The sample includes all non desk-rejected papers. "# of Days to Submit Review" is the number of days between a referee accepting an invitation and submitting a report, conditional on the report not being late. "# Reminders Sent After Due Date" is conditional on a review being submitted late. "Editor is experienced" is a dummy variable that takes a value of 1 if an editor's years of experience is greater than the average value among all editors in the sample, and 0 otherwise. Regressions control for year invited, and referee experience. Standard errors are clustered by manuscript. * p<0.10, ** p<0.05, *** p<0.01.

## Table A3: Report Content

| | Referee Report Wordcount | | Referee Report Sentiment | | Letter to Editor Wordcount | | Letter to Editor Sentiment | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| referee is female | 83.495* | 55.323 | -0.008** | -0.012* | 0.846 | 40.786 | -0.001 | -0.005 |
| | (47.648) | (70.243) | (0.004) | (0.006) | (29.293) | (43.608) | (0.011) | (0.018) |
| editor and ref are female | | 86.953 | | 0.006 | | -62.855 | | 0.015 |
| | | (111.579) | | (0.009) | | (63.785) | | (0.024) |
| ref is female x editor is experienced | | 11.123 | | 0.008 | | -73.343 | | 0.001 |
| | | (106.186) | | (0.009) | | (73.089) | | (0.027) |
| Observations | 1750 | 1750 | 1750 | 1750 | 1346 | 1346 | 1346 | 1346 |
| $R^2$ | 0.501 | 0.501 | 0.564 | 0.565 | 0.445 | 0.446 | 0.449 | 0.449 |
| controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Manu FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Mean | 1079.226 | 1079.226 | .077 | .077 | 262.969 | 262.969 | .142 | .142 |

Notes: Observations are unique at the referee-manuscript level. The sample includes all non desk-rejected papers. Both "Referee Report Sentiment" and "Letter to Editor Sentiment" are measures that take values of [-1,1] and are calculated from a rule-based sentiment analysis of the respective texts. "Editor is experienced" is a dummy variable that takes a value of 1 if an editor's years of experience is greater than the average value among all editors in the sample, and 0 otherwise. Regressions control for year invited, and referee experience. Standard errors are clustered by manuscript. * p<0.10, ** p<0.05, *** p<0.01.

Table A4: Report Content - Alternative Sentiment Measure

| | Referee Report Sentiment | | Letter to Editor Sentiment | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| referee is female | 0.021 | -0.000 | -0.013 | -0.048 |
| | (0.033) | (0.045) | (0.038) | (0.061) |
| editor and ref are female | | -0.003 | | 0.064 |
| | | (0.075) | | (0.084) |
| ref is female x editor is experienced | | 0.075 | | 0.053 |
| | | (0.075) | | (0.098) |
| Observations | 1750 | 1750 | 1346 | 1346 |
| $R^2$ | 0.637 | 0.637 | 0.527 | 0.528 |
| controls | Yes | Yes | Yes | Yes |
| Manu FE | Yes | Yes | Yes | Yes |
| Mean | .818 | .818 | .752 | .752 |

Notes: Observations are unique at the referee-manuscript level. The sample includes all non desk-rejected papers, but is limited to referees for whom we have referee reports (i.e., the sample from Table 5, Columns (1)-(4)). Both "Referee Report Sentiment" and "Letter to Editor Sentiment" are measures that take values of [-1,1] and are calculated from an rule-based sentiment analysis of the respective texts. The method in this table uses a different dictionary of positive and negative words than the measure in our main table. "Editor is experienced" is a dummy variable that takes a value of 1 if an editor's years of experience is greater than the average value among all editors in the sample, and 0 otherwise. Regressions control for year invited, and referee experience. Standard errors are clustered by manuscript. * p<0.10, ** p<0.05, *** p<0.01.

# References


Abrevaya, J., and D. 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.

Babcock, L., B. Peyser, L. Vesterlund, and L. Weingart (2022): *The No Club: Putting a Stop to Women's Dead-end Work*. Simon and Schuster.

Ginther, D. K., and S. Kahn (2021): "Women in academic economics: Have we made progress?," in *AEA Papers and Proceedings*, vol. 111, pp. 138–142. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Hengel, E. (2022): "Are women held to higher standards? Evidence from peer review," *The Economic Journal*.

Hengel, E., and E. Moon (2020): "Gender and equality at top economics journals," .

Kuminoff, N. V., K. E. Ciaramello, H. M. Dooley, M. D. Heintzelman, N. Khanna, L.-R. Kosnik, L. Y. Lewis, and E. Trimble (2023): "New Evidence on Diversity in Environmental and Resource Economics," *Review of Environmental Economics and Policy*, 17(1), 000–000.