INTERNATIONAL ECONOMIC REVIEW
Vol. 00, No. 0, February 2021 DOI: 10.1111/iere.12503

GENDER BIAS IN OPINION AGGREGATION*

By Friederike Mengel

University of Essex, U.K. ; Lund University, Sweden

Gender biases have been documented in many areas including hiring, promotion, or performance evaluations. Many of these decisions are made by committees. We experimentally investigate whether committee deliberation contributes to gender biases. In our experiments, participants perform a real effort task and then rate the task performance of other participants. Across treatments we vary the extent of deliberation possible. We find that deliberation increases gender biases. We explore several mechanisms and test two interventions. Randomizing the order of speaking does not reduce gender bias, but an information intervention where raters are informed of gender bias in prior sessions does.

1. INTRODUCTION

Persistent gender earnings and promotion gaps have attracted much attention in research and policy debates in recent years; see Goldin and Rouse (2000), Black et al. (2008), or Sandberg (2013) among many others.[1] Indeed, a large body of empirical evidence has documented gender biases in decisions, such as hiring (Neumark et al., 1996; Goldin and Rouse, 2000), promotion (Booth et al., 2003; Ginther and Khan, 2004; Bagues and Esteve-Volart, 2010; Bagues et al., 2017), the allocation of venture capital investments (Malmstrom et al., 2018), or performance evaluations (Bohnet et al., 2016; Sandberg, 2018). One thing that is common to all of these decisions is that they involve deliberation by committee members.

In this article, we aim to understand whether committee deliberation contributes to gender biases. Doing so requires (i) measuring pre- and postcommunication beliefs and (ii) an experimental variation of the amount of deliberation allowed. Conducting a lab experiment enables us to create such a design. In all treatments of our experiment, participants perform a real effort task, where performance evaluation is subjective. They then rate the performance of nine other (anonymous) participants. Our treatments systematically vary two features that distinguish committee decision making from individual decisions: (i) the amount of deliberation possible and (ii) the fact that there are incentives to reach an agreement in the committee.

*Manuscript received April 2020; revised September 2020.

I thank Sonia Bhalotra, Steven Bosworth, Irma Clots Figueras, Chiara Franzoni, Ilyana Kuziemko, Michael Naef, Wieland Mueller, Johanna Rickne, Daniel Sgroi, Camille Terrier, and seminar participants in Amsterdam, Birmingham, Essex, Innsbruck, King's College London (Political Economy workshop), Konstanz, LMU Munich, Lund (Microseminar and Arne Rhyde Conference on Learning and Evolution), Nice (ASFEE 2018), Nuremberg (Conference on Gender and the Labour Market 2018), NYU Abu Dhabi, Reading (Behavioural Econ Workshop), and Toulouse for helpful comments and Sara Godoy, Mihail Morosan, Rafael Brancu, and Flavio Lomaski Torrez for excellent research assistance. Financial support by the European Research Council (ERC Starting Grant 805017-DYNNET) is gratefully acknowledged. Please address correspondence to: Friederike Mengel, Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, U.K.; Department of Economics, Lund University, SE-220 07 Lund, Sweden. E-mail: *fr.mengel@gmail.com*

[1] Academia is no exception. Although in some fields of academia gender promotion and earnings gaps have converged, this is not true in others. Economics is one of the fields where promotion and earnings gaps are particularly persistent and cannot be easily explained. According to Ceci et al. (2014), earnings gaps among U.S. full professors in Economics are even higher in 2010 than they were in 1995 with female full professors earning less than 75% of their male counterparts.

Our main treatments involve open committee deliberation. Participants first rate task performance by other experimental participants. They can then deliberate via a chat window for three minutes with two other committee members. After the chat, they see the ratings of the other committee members and submit a revised rating. Committees are given an incentive payment if all three revised ratings agree.

We then systematically shut down different aspects of committee deliberation. First, we remove the chat and with it the possibility to persuade or convince others. Next, we remove the possibility to exchange information (to see others' ratings), and last we also remove the incentive payment, leading us to a situation of individual decisions where participants do not have to think about what others might do. All treatments exist in a variation where gender identity is revealed (**G** variation) and one where it is not. Comparing the gender-blind and nonblind treatments identifies the gender bias in the different conditions.

We find strong and highly statistically significant gender biases under open committee deliberation. After deliberation, 60% of ratings received by men are revised upward compared to only 25% of ratings received by women. As a consequence women are ranked on average three positions lower after deliberation.[2] Shutting down open deliberation and information exchange removes the gender bias at least as long as there are still incentives to agree. However, doing so might not be desirable viewed from other perspectives. It might not lead to optimal decisions, for example, if decision makers hold different information about candidates. It might also not be feasible for legal reasons or when communication among committee members cannot practically be prevented.

We hence tested two further interventions both designed to reduce gender bias in the presence of open deliberation. The first intervention randomized the order of speaking in the committee. This intervention was unsuccessful and in fact produced weakly larger gender biases compared to our baseline open-deliberation treatment. The second intervention we tested is an information intervention, where participants are made aware of gender bias in previous sessions prior to entering their ratings. Similar interventions have sometimes been shown to be successful in noncommittee decision making (Boring and Philippe, 2017; Pope et al., 2018). We also find that this intervention is successful. There is no gender bias in this treatment.

These results carry potentially actionable policy consequences. Our interventions have shown that care must be taken when designing rules for committee deliberation. Changes designed to reduce bias, such as randomizing the order of speaking in a committee, can have unintended consequences and in our case led to very strong gender bias. On the other hand, our information intervention was successful and did not lead to gender bias (neither against men nor women). We also did not find evidence that this intervention would lead to greater polarization of opinions.

Our results contribute to two strands of literature that we will review in detail below (Section 2): (i) literature on information aggregation in groups and (ii) literature on gender bias. With respect to the former our results point to the importance of institutional detail when "truth" is subjective. Under minimal communication there is no gender bias and committees arguably reach a more objective judgement. Under open communication, by contrast, the group is more biased than the sum of individual ratings would suggest. With respect to literature on gender bias, we highlight the importance of studying the role of committee decision making in many of the areas where gender biases have been identified. To our knowledge, our article is the first to identify the role of committee deliberation for gender biases.

The article is organized as follows. Section 2 discusses literature on information aggregation and gender biases in more detail and points out how our article contributes to each. Section 3 describes the experimental design and procedures. Section 4 contains the main results. Section 5 discusses the results from two different "interventions" designed to reduce gender

[2] We also conduct a sentiment analysis (Thelwall et al., 2010), which reveals that chats contain more positive statements when a summary written by a man is rated compared to when a woman's summary is rated.

bias under open deliberation and Section 6 concludes. Experimental instructions, additional tables, figures, and information about the sample can be found in a series of Appendices.

2. RELATED LITERATURE

Our research contributes to two so far largely disjoint strands of research: (i) research on deliberation and information aggregation and (ii) research on gender biases. We will review these in turn.

2.1. *Literature on Deliberation and Opinion Aggregation.* The literature on deliberation and opinion aggregation is extensive. Most of it deals with the aggregation of dispersedly held private information in contexts where there is an underlying truth value that can be learned. One of the striking findings from this literature is that when groups try to learn an objective fact (say, e.g., a probability), the group often does better than one would expect by simply averaging each group member's prior. This fact has become known as the "wisdom of the crowd" (Surowiecki, 2004) and has recently been demonstrated in the context of experimental economics by Della Vigna and Pope (2018). Groups have also been shown to learn faster and make "more rational" decisions in strategic situations compared to individuals (Cooper and Kagel, 2005; Kocher and Sutter, 2005; Cooper and Kagel, 2016).[3] This literature suggests that the process of deliberation could reduce biases and lead to better decision making.[4] However, there are also results showing that certain forms of structured communication can lead to information cascades where people fail to learn the truth, but instead herd on wrong information (Anderson and Holt, 1997). The difficulty of reaching a consensus and learning the truth has also been pointed out in recent experimental literature on opinion dynamics in social networks (Corrazini et al., 2012; Brandts et al., 2015; Chandrasekhar et al., 2015; Grimm and Mengel, 2020). Other reasons for failing to learn an objective truth are social pressure, as in the classic experiment by Asch (1995), or a desire to be perceived favorably by other group members (Isenberg, 1986). Researchers have also pointed out the dangers of "groupthink," that is, situations where groups reach an often evidently wrong consensus (Janis, 1972; Turner and Pratkanis, 1998; Benabou, 2013).[5]

Closer to our research is literature dealing with the aggregation of information of a more subjective nature, such as risk assessments, political beliefs, or preferences. This literature is, on balance, more pessimistic about consensus. A number of results have demonstrated polarization of opinions, that is, cases where groups fail to reach a consensus and where each group members' opinions are more "extreme" than they were before deliberation (Sunstein, 2000; Baldassari and Bearman, 2007). Such effects have been documented with respect to altruism (Cason and Mui, 1997) or political opinions (Baldassari and Bearman, 2007). Cason and Mui (1997), for example, have found evidence of group polarization in the context of dictator game giving (see also Luhan et al., 2009). A number of researchers have studied aggregation of risk preferences (Shupp and Williams, 2008; Masclet et al., 2009; Casari and Zhang, 2012; Ozdemir, 2018). Among those Ozdemir (2018)'s work is probably the closest as it asks whether group conformity can play a role in explaining why fewer female leaders are elected. She asks "leaders" to make decisions under risk on behalf of a group and has the group elect the leader. A key difference to our work is that aggregation in her paper happens via voting and not via deliberation. Ambrus et al. (2015) study preference aggregation and they do elicit individual opinions before deliberation. They find that positions close to the median of

[3] Interestingly Cooper and Kagel (2016) find that women are less likely to give advice compared to men in team play. In terms of our measures of participation in communication (see Subsection 4.4.4), we do not find such an effect.

[4] Philosophers such as John Rawls have pointed out deliberation as a means to reduce biases. Rawls writes, "In everyday life the exchange of opinion with others checks our partiality and widens our perspective..." (Rawls, 1971).

[5] Both Janis (1972) and Benabou (2013) provide a collection of illustrative examples where failure was attributed to "groupthink," such as, for example, the Columbia space shuttle accident, financial market hubris, the Bay of Pigs invasion, or the Cuban missile crisis.

a group are most influential in the group consensus. This is true, on average, in our study as well. We do find, however, that upward biased opinions have a strong positive impact on group consensus if and only if the rated person is male. There is also some literature on deliberation and aggregation of heterogenous preferences in the political economy literature. Goeree and Yariv (2011) study collective deliberations when information aggregation happens via different voting schemes and when voters have differing preferences over outcomes. They find that deliberation improves efficiency. There are some key differences between these papers and our research. First, except for Ozdemir (2018), they do not study gender bias. Second, if they allow for deliberation at all, they only consider open forms of communication. Third, they do not study aggregation of performance evaluations. Experimental literature on aggregation of subjective performance assessments is scarce and to our knowledge ours is the first article studying biases when such subjective assessments are aggregated. This is one of the main features of how our article differs from this literature.

2.2. *Literature on Gender Biases.* Our study also contributes to a large and diverse literature on gender biases in performance evaluations (Goldin and Rouse, 2000; Bagues and Esteve-Volart, 2010; Bohnet et al., 2016; Bagues et al., 2017; Boring, 2017; Mengel et al., 2019; Malmstrom et al., 2018).[6] Particularly relevant among those studies to our case is maybe Malmstrom et al. (2018), as they study committee deliberation. They analyzed communication among venture capital investors and find that they allocate less to female entrepreneurs based on common perceptions about women being less risk taking or less ambitious than men. Their study does not, however, contain an experimental variation of deliberation and hence cannot identify the effect of deliberation on gender biases. Coffman et al. (2020) study the effect of gender stereotypes on deliberation. They find that people put more weight on others as well as themselves in deliberation if there is a strong stereotype associated with task performance by their (others) gender. Schwarz and Vesterlund (2020) find that the gender composition of the committee matters in a lab experiment. Prior to us psychologists have used lab experiments to understand individual gender biases in performance evaluations. Swim et al. (1989), for example, review literature on a classic experiment where students had to rate identical articles or poems that were supposedly written by a Joan McKay or John McKay. They conclude that the evidence on whether ratings are biased is mixed and effect sizes quite small. Many of these experiments involve deception and most are not incentivized (see also MacNell et al., 2015). Krawczyk and Smyk (2016) conduct an incentivized and deception-free lab experiment, where they let students rate scientific articles written by male and female economists. They find that if gender is revealed, articles written by female authors are rated worse relative to articles written by male authors. Our contribution to this literature is to show how committee deliberation affects gender biases.[7] Next, we will describe the design of our experiment.

3. DESIGN OF THE EXPERIMENT

Our experiment consists of eight main treatments as well as some interventions. In all treatments participants first completed a task and then rated tasks by other participants. We start by describing the task.

3.1. *The Task.* We were looking for a task that (i) participants are familiar with, (ii) where performance will differ across participants (with as few identical solutions as possible), and (iii) where importantly performance ratings are subjective. We decided to let participants summarize a news article conveying its key message in at most 1,000 characters. The

[6] See Altonji and Blank (1999) for a survey of earlier literature on gender bias in the labor market.

[7] Gender bias in deliberations has also been studied in law and psychology (see, e.g., Marder, 1987). In this literature, the focus was often on representativeness of a jury and whether or not female jury members get equal weight in the ultimate decision of the group.

article can be found in Figure B.1 in Online Appendix B together with samples of the best-rated and worst-rated summaries.

Gender stereotypes about such a task are ambiguous. Women are usually thought to be better with verbal skills and language (Plante et al., 2009), men with technical topics (such as the subject matter of the article), and men are usually thought to be better at "summarizing" (Holmes and Meyerhoff, 2003). We also asked respondents in an online survey ($n = 136$) about whether they thought that men or women performed better in the task. Although the modal answer is "about the same" (47% of answers), substantial minorities believe that women performed "a little better" (32%) or that men performed "a little better" (11%). Survey responses also point to ambiguous stereotypes when participants were asked what the most important skills are to complete the task successfully with the most prominent answers including both male-stereotyped (intelligence, logical reasoning) as well as female-stereotyped (language skills, good organization) skills. See Online Appendix D for more details on this survey.

Each treatment exists in a "gender blind" version, where gender identity is not revealed, and in a "gender" (**G**) version, where gender identity is revealed.

3.2. *Revealing Gender Identity.* Our aim was to reveal gender identity in a way that (i) seems natural to participants, (ii) is credible, and (iii) does not trigger suspicion that the experiment is about gender. We proceeded as follows: At the start of the experiment participants were asked to enter some basic demographic information (age, gender, field of study, etc.). This happened in all treatments. Afterward, in the **G** versions only, participants were informed on the screen that they were assigned an avatar, which was shown to them on the screen. They were also informed that "*All women have been assigned a female 'avatar' and all men a male 'avatar.' Other than that the pictures have no connection to the information given by you*." Hence they were informed that behind a female (male) avatar is always a female (male) participant, but this information was framed in the context of an assurance of their anonymity. At the time of filling in their demographics, participants did not know yet that there would be avatars in the experiment.[8] We used 24 different female and male avatars to reduce the risk that particular facial features might trigger responses by others and we check for differences across avatars.[9] We also investigate in detail in Section D.1 whether this design choice induced priming effects. Finally, evidence from an open question ("What do you think this experiment was about?") at the end of the postexperimental questionnaire suggests that this design choice was successful in achieving goal (iii), that is, participants did *not* perceive this as an experiment about gender.

We now describe the three communication variations, which complete the 3×2 design.

3.3. *Communication Variations.* In the treatments **NO** and **NOG,** there are no communication possibilities. Participants complete the task and then sequentially rate nine tasks from other participants (in random order) without any information on others' ratings. Tasks are rated on a scale from 0–10 where 0 is worst and 10 is best. Participants are paid (i) the average rating their task received in pounds (between 0 and 10 GBP) and (ii) a show up fee of 3 GBP. Hence, although the quality of the summary was incentivized, the ratings given to other summaries were not. There could potentially be two ways to incentivize ratings. First, one could pay for the quality of ratings by comparing them to an objective measure of quality of a summary. Such a measure is, however, not available (see subsection "*The Task*"). A second possibility is to pay for how close ratings are to the average rating of others either inside or

[8] The instructions do not make any mention of avatars, so participants learn that avatars exist only after they have filled in the demographic information. (See the instructions in Online Appendix A.)

[9] Figure F.2 in Online Appendix F shows the average ratings received by different avatars and illustrates that they are very similar with no notable outliers. We also used the same avatars in a different study and found no difference across avatars in terms of the altruism directed toward them or the willingness of people to "network" with them (Mengel, 2020).

outside the experiment. Whereas the former possibility is ruled out by our desire to have a task where performance is subjective, the latter possibility would turn ratings into a beauty contest, where participants do not express their own opinion but try to match that of others. As a consequence, for the experimenter it would become impossible to distinguish gender bias from beliefs about gender bias by others. Because of these concerns, we decided not to give incentive payments for ratings in the **NO** treatments. Committee deliberation adds two qualitatively different elements.

First, there are incentives to agree. Committee members typically need to come to a decision at the end of deliberations, requiring some degree of compromise or agreement. We conducted two treatments (**BASE-IAG** and **BASEG-IAG**) that coincide with the **NO** treatments, but where we provided participants with an incentive to agree. Specifically, at the beginning of the experiment, they were matched with two other participants ("committee members") and paid in each round an additional 3 GBP if the ratings of all three group members coincided in that round.[10] Note that, although in the **BASE-IAG** treatments, participants have to think how others might rate in order to obtain the incentive payment, there is still no communication among the committee members. Compared to the **NO** treatments, treatments **BASE-IAG** hence identify the impact of having incentives to agree.

Second, there is communication. Committee members talk, exchange information on each others views, try to persuade others, etc. We introduce communication in two steps. In treatments **MIN** and **MING** there are minimal communication possibilities. Participants were randomly allocated in groups of three. They then rated nine tasks as in the **NO-IAG** condition, but never got to rate tasks of members of their own group. The difference to the **NO-IAG** condition is that, for each of the nine tasks, after submitting their initial rating, participants got to see the ratings of all three committee members (including their own) and were invited to submit a revised rating. Communication here is minimal as participants can only see their group members' ratings without having any chance to convince or persuade others. Participants were paid as before, including the incentive to agree. Specifically, in each round, they were paid an additional 3 GBP if the *revised* ratings of all three group members coincided.

Treatments **OPEN** and **OPENG** offer open communication possibilities. After submitting their initial rating, participants in these treatments got to a chat screen, where they could chat via typed messages with other group members for three minutes. Afterward, they were shown the initial ratings of all three group members (including their own) and were invited to submit a revised rating. Other than the addition of the chat, these treatments were identical to the minimal communication variations. Treatments **OPEN** and **OPENG** are our main treatments of interest as they mimic the full committee decision process.

Across all deliberation variations, gender bias is identified by comparing ratings across the "gender blind" versus "gender revealed" (**G**) versions of each variation.

3.4. *Intervention Treatments.* We also conducted two intervention treatments, where we made changes to the **OPEN** condition with the potential to reduce gender bias. The first intervention was to randomize the order of speaking in the open-deliberation process (**OPEN-RAND**). The second was to provide participants with information that gender bias had been identified in previous sessions (**OPENG-INFO**). More details also on the motivation behind these treatments are discussed in Section 5.

3.5. *Sample, Other Details, and Randomization Checks.* Table 1 summarizes some basic information about treatments. In total, 682 people participated in our lab experiment conducted

[10] There are many other possibilities to incentivize agreement. We could have chosen to pay whenever a majority agrees or paid in (inverse) proportion to the standard deviation or a similar rule. We chose to pay if all group members agree as (i) it is a simple rule and (ii) it avoids coordination problems relating to whose ratings to align.

TABLE 1
SUMMARY OF TREATMENTS, NUMBER OF PARTICIPANTS (N), AND PERCENTAGE OF MALE PARTICIPANTS AS WELL AS PRE- AND POSTCOMMUNICATION MEAN RATINGS FOR WOMEN (W) AND MEN (M)

					Precomm		Postcomm	
	N	% Male		N	W	M	W	M
NO: no communication; no incentives to agree	120	49%	blind	60	6.10	5.68	–	–
			not blind (**G**)	60	6.03	6.25	–	–
NO-IAG: no communication; with incentives to agree	126	48%	blind	63	6.33	6.66	–	–
			not blind (**G**)	60	6.03	6.25	–	–
MIN: minimal communication; with incentives to agree	127	47%	blind	66	6.09	6.34	6.15	6.38
			not blind (**G**)	63	6.45	6.57	6.50	6.68
OPEN: chats; with incentives to agree	126	57%	blind	60	6.22	6.12	6.22	6.07
			not blind (**G**)	66	6.55	6.52	6.50	6.81
OPEN-RAND: chats; random speaking order; with incentives to agree	120	50%	blind	60	5.89	5.89	5.88	5.87
			not blind (**G**)	60	5.13	5.96	5.06	6.20
OPEN-INFO: chats; information provided; with incentives to agree	63	62%	blind	–	–	–	–	–
			not blind (**G**)	63	6.53	6.53	6.53	6.47

at Essex Lab at the University of Essex.[11] Online Appendix C contains further details on the demographics of our sample in terms of age, gender, nationality, and occupation. We provide two types of balancing checks in Online Appendix C. First, we compare all treatments to treatment **NO** (Table C.1). Here, we find that participants in **OPEN-RAND** are somewhat older on average. Furthermore, in both **OPEN-RAND** treatments, there are somewhat fewer students and fewer participants from outside the EU. Our second randomization check is maybe more crucial. Here, we compare the gender-blind and nonblind versions of all deliberation variations (Table C.2). Of the 30 comparisons we make, only one is statistically significant at the 5% level, which is that there seem to be somewhat more participants from outside the EU in **MING** compared to **MIN**. Broadly, our treatments seem balanced with respect to participant characteristics. On average, sessions without communication lasted about 55 minutes and those with communication lasted about 80 minutes. Payments ranged between GBP 5.44 and GBP 15.33 with a mean of GBP 10.20. Ethical approval was obtained by the Social Sciences Faculty Ethics subcommittee at the University of Essex.

3.6. *Online Surveys.* We also conducted three types of online surveys. In December 2014, we fielded a survey with a professional online survey provider, who maintains a sample that is representative of the U.K. adult population. We asked a subsample of 439 independent raters to rate the summaries written by our participants in treatments **NO** and **NOG**. This was done in order to understand whether ratings can broadly be reproduced by independent raters. Four hundred thirty-nine participants participated in the survey, though some (less than 5%) dropped out midway through the survey. Online Appendix D.1 contains more information on this survey. In September 2020, we ran a similar online experiment. We now had independent raters rate the summaries written in **OPENG**. Each summary was rated exactly nine times (by nine different raters), just as in our lab experiment. A second type of survey was fielded in

[11] Sessions were conducted in 2014 (basic treatments), 2018 (intervention treatments), and 2020 (additional sessions for all treatments after complaints by reviewers about low power). Results did not change qualitatively after increasing power in January 2020.

TABLE 2
FINAL RATINGS FROM OPEN-DELIBERATION TREATMENTS REGRESSED ON AUTHOR GENDER, A DUMMY INDICATING WHETHER THE TREATMENT IS ONE WHERE GENDER IDENTITY IS REVEALED, AND THE INTERACTION OF THE LATTER WITH AUTHOR GENDER

	Open Deliberation					
	Baseline: **OPEN** treatment				Baseline: Online Study	
	(1)	(2)	(3)	(4)	(5)	(6)
male (β_1)	−0.151	−0.182	−0.308**	−0.279***	−0.147	−0.155
	(0.093)	(0.132)	(0.135)	(0.102)	(0.120)	(0.105)
δ_G (β_2)	0.275	−0.042	−0.121	0.098	−0.091	−0.091
	(0.167)	(0.236)	(0.241)	(0.201)	(0.129)	(0.130)
$\delta_G \times$male (β_3)	0.461*	0.534**	0.726***	0.589***	0.457**	0.457**
	(0.239)	(0.252)	(0.276)	(0.188)	(0.178)	(0.178)
Constant	6.225***	7.071***	7.194***	6.417***	6.592***	6.761***
	(0.0801)	(0.483)	(0.491)	(0.376)	(0.0734)	(0.424)
Drop Constant Raters	–	–	✓	✓	–	–
Session FE	–	✓	✓	✓	✓	✓
Demographics	–	✓	✓	✓	–	✓
p-values						
Test $\beta_1 + \beta_3 = 0$	0.1625	0.1262	0.1004	0.0499	0.1656	0.1722
Observations	1,215	1,215	1,161	946	1,350	1,350
R-squared	0.018	0.055	0.067	0.099	0.005	0.044

NOTES: Except for column (1) we also control for author demographics (age, whether the author is a student, and nationality fixed effects) and session fixed effects. In column (3) we additionally drop all raters who enter the same rating for all summaries (see footnote). Column (4) additionally drops all ratings above the 95th and below the 5th percentile. Columns (5) and (6) use the ratings from an online study, where independent raters rated the summaries produced in **OPENG** in a gender-blind manner, as a baseline (instead of ratings from **OPEN**). Robust standard errors clustered at the rater level.
Robust standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

September 2018. In this survey, we asked 136 respondents about stereotypes associated with the task. A third type of survey was fielded in January 2019, where we asked 117 participants to classify messages from treatment **OPEN-RAND** according to whether they are "strong" or "weak" statements. This survey will be discussed in Subsection 5.1. Details about all online studies can be found in Online Appendix D.

4. MAIN RESULTS

This section contains our main results. We first focus on the decisions made in the full committee setting, that is, treatments **OPEN** and **OPENG**. Whereas in Subsection 4.1 we focus on final ratings, Subsection 4.2 then focuses on the effect of deliberation, that is, the change between pre- and postcommunication ratings. Subsection 4.3 discusses different mechanisms including what happens when we strip away communication possibilities and incentives to agree. Intervention treatments are discussed in Section 5.

4.1. *Committee Decisions: Final Ratings.* We ask how gender affects final ratings with open deliberation. Table 2 shows results of estimating the following equation:

$$\texttt{Rating}_{ij} = \alpha + \beta_1 \texttt{male}_i + \beta_2 \delta_G + \beta_3 (\delta_G \times \texttt{male}_i) + X_{ij} + \epsilon_{ij}, \tag{1}$$

where the rating author i obtains from rater j ($\texttt{Rating}_{ij}$) is regressed on author gender ($\texttt{male}_i$), a dummy δ_G indicating whether the treatment is one where gender identity is revealed, and the interaction of the latter with author gender. X_{ij} is a vector of controls including author characteristics, session fixed effects, and (in specifications reported in Online Appendix Table E.4) committee fixed effects. Standard errors are clustered at the rater

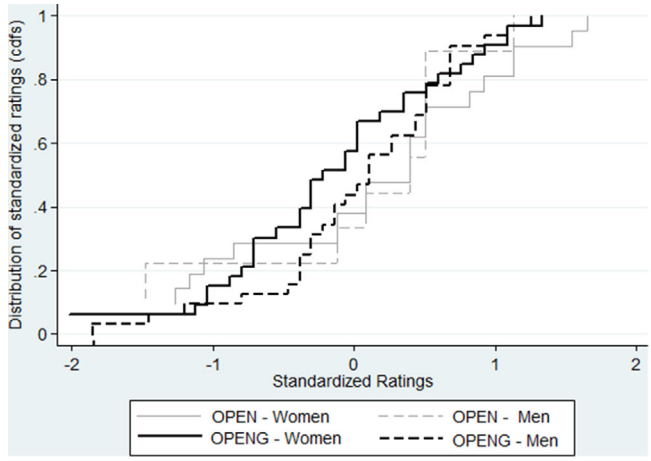


FIGURE 1

GENDER BIAS WITH COMMUNICATION: DISTRIBUTION OF STANDARDIZED RATINGS IN **OPEN** AND **OPENG** FOR BOTH MEN AND WOMEN [COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

level.[12] In some specifications, we drop "constant raters," which are raters who enter the same rating for all summaries.[13] The main focus of the article is on the coefficient β_3, which measures gender bias. β_3 identifies gender bias whenever there are no performance differences between men and women in the gender-blind versus gender-revealed (**G**) conditions. There is nothing in the data to suggest that this identifying assumption would not be satisfied (see also balancing checks reported in Online Appendix C). To be sure, we also conducted an online experiment where independent raters rated the summaries produced in **OPENG** in a gender-blind manner. This allows us to literally hold performance constant across the gender-blind and gender-revealed conditions.

Table 2 shows the results of running regression (1) on final ratings in the open-deliberation treatments. Columns (1)–(4) show the results from the lab experiment, that is, of treatments **OPEN** and **OPENG**. Under open deliberation, gender bias (β_3) is substantial and highly statistically significant. The relative position of women worsens when gender identity is revealed with men gaining about 0.45–0.7 rating points over women. In columns (5) and (6) we use the ratings from the online study where independent raters rated the summaries produced in **OPENG** in a gender-blind manner as a baseline instead. Also in this case there is a highly statistically significant gender bias with $\beta_3 = 0.457$.[14] It is also worthwhile to note that in the gender-blind conditions, summaries produced by women receive somewhat higher ratings on average, though the difference is not statistically significant in all specifications.

Figure 1(a) shows the distribution of standardized ratings in **OPEN** and **OPENG** for both men and women. The figure shows that, although there is no difference in the rating

[12] Standard errors hence account for the fact that ratings can be correlated across the different summaries rated by each rater. However, there are other ways to cluster standard errors, too. In particular, in the communication treatments it seems to make sense to also account for the fact that raters in the same committee can influence each other. In Online Appendix E, we show regressions that include committee fixed effects and where standard errors are clustered at either the author or the rater level (Table E.4). Those regressions show broadly the same results.

[13] The number of constant raters is small, specifically two in **NO**, two in **OPEN**, three in **MING**, four in **OPENG,** and one each in treatments **OPENG-RAND** and **OPEN-INFO** (see Section 5 for results from these treatments). There are no constant raters in any of the other treatments. Constant raters always enter the same rating for each summary and in 10 out of the 11 cases identified this is the maximum rating of 10. Because of this even few constant raters can affect effect sizes substantially.

[14] The pairwise correlation between the ratings in the independent online experiment and treatment **OPENG** is 0.4019***. This shows that ratings are meaningful. More details on the online study can be found in Online Appendix D.1.

TABLE 3
OLS REGRESSION OF DIFFERENCE BETWEEN POST AND PRECOMMUNICATION RATINGS (COLUMNS (1)–(3)) AND RANK BASED ON AVERAGE RATINGS (COLUMNS (4) AND (5)) ON AUTHOR GENDER, A DUMMY INDICATING WHETHER THE TREATMENT IS ONE WHERE GENDER IDENTITY IS REVEALED, AND THE INTERACTION OF THE LATTER WITH AUTHOR GENDER

	Open Deliberation				
	Rating Differences			Rank Differences	
	(1)	(2)	(3)	(4)	(5)
male (β_1)	−0.054	−0.079	−0.100	1.493*	1.061
	(0.096)	(0.112)	(0.117)	(0.648)	(0.903)
δ_G (β_2)	−0.055	−0.067	−0.057	3.722***	4.091***
	(0.089)	(0.091)	(0.093)	(0.760)	(1.002)
$\delta_G \times$male (β_3)	0.393***	0.409***	0.448***	−8.189***	−7.205***
	(0.116)	(0.129)	(0.140)	(0.793)	(1.008)
Constant	0.003	0.283	0.288	−0.597	−1.726
	(0.067)	(0.275)	(0.280)	(0.545)	(4.440)
Drop Constant Raters	–	–	✓	–	–
Session FE	–	✓	✓	–	✓
Demographics	–	✓	✓	–	✓
Observations	1,215	1,215	1,161	135	135
R-squared	0.016	0.028	0.030	0.108	0.173

NOTES: Author demographics are age, whether the author is a student, and a nationality indicator. Standard errors are clustered at the rater level for columns (1)–(3) and at session level in columns (4) and (5). Open-deliberation treatments.
Robust standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

distribution of men and women in **OPEN** (gray lines), the distribution of average ratings received by men first order stochastically dominates that of women in **OPENG** (black lines), except for a region near the top of the distribution. The median final rating of women is about 22% of a standard deviation lower than that of the median man. We next study the effect of deliberation and then turn to a discussion of implications and the economic significance of the gender bias identified in this treatment.

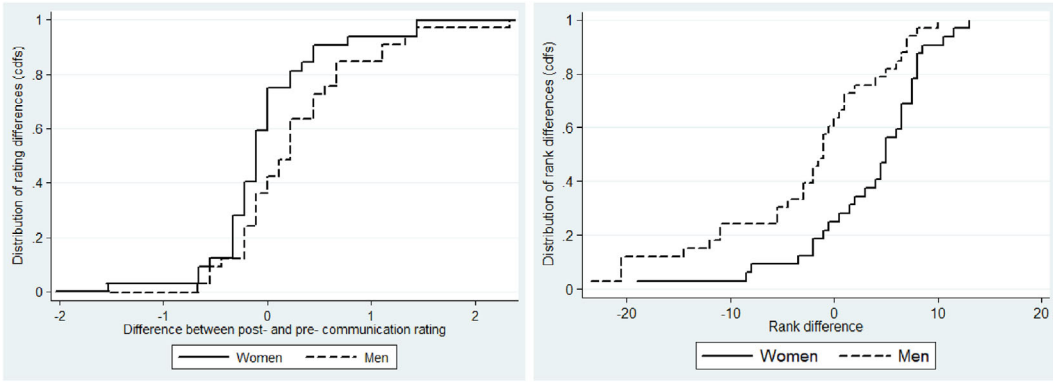
4.2. *Effect of Committee Deliberation.* In order to study the effect of deliberation on ratings we focus on differences between pre- and postcommunication ratings received by women and men. We analyze rating differences induced by communication by estimating the following equation:

$$\Delta_{ij} = \alpha + \beta_1 \texttt{male}_i + \beta_2 \delta_G + \beta_3 (\delta_G \times \texttt{male}_i) + X_{ij} + \epsilon_{ij}. \tag{2}$$

The endogenous variable Δ_{ij} here is the difference between pre- and postcommunication ratings received by author i from rater j and the exogenous variables are the same as in Equation (1). Coefficients α and $\alpha + \beta_1$, respectively, show the effect of deliberation on women's and men's ratings in the gender-blind condition, whereas $\alpha + \beta_2$ and $(\alpha + \beta_1 + \beta_2 + \beta_3)$ show the effect of deliberation on women's and men's ratings, respectively, in the treatments where gender identity is known. As before, the differential effect of revealing gender for men and women identifies the gender bias and shows up in the regression with coefficient β_3.

Table 3 (columns (1)–(3)) shows the results of running regression (2) for the open-deliberation treatments. The gender bias found above is clearly reflected in what happens between pre- and postcommunication ratings. Men's ratings increase by about 0.4 points more postcommunication than women's when gender identity is revealed compared to when it is not. This gender bias is highly statistically significant in all specifications.

Figure 2 (Panel a) shows the distribution of these differences for treatment **OPENG**. The distribution of pre- and postcommunication differences for men first order stochastically



(a) Ratings (b) Rank

FIGURE 2

DISTRIBUTION (CDF) OF DIFFERENCE IN POST- AND PRECOMMUNICATION AVERAGE RATINGS (PANEL A) AND RANK (PANEL B) OBTAINED BY WOMEN AND MEN IN TREATMENT **OPENG** [COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

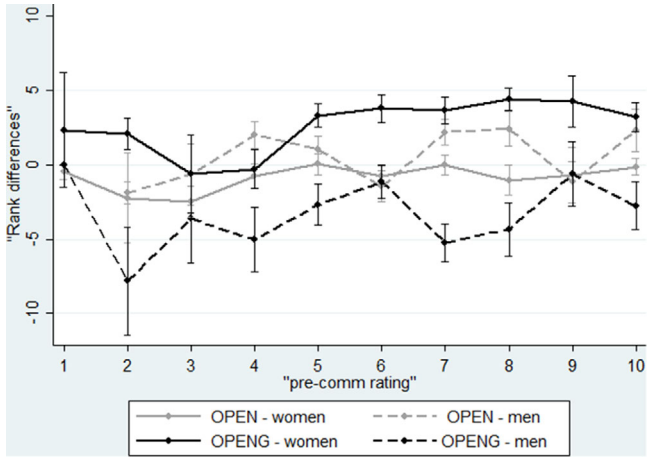
dominates that of women. Around 60% of men's ratings are increased in the communication process compared to only $\approx 25\%$ of women's, whose received ratings tend to decrease or stay the same postcommunication. Hence, open deliberation leads to gender bias and it is the process of deliberation itself that induces the gender bias.

The implications of this bias can indeed be severe. If these ratings were at the basis of a pairwise competition for a job, a promotion, or similar, then our findings in treatment **OPENG** imply that a woman who is rated higher than her male competitor precommunication has an $\approx 40\%$ chance to be rated lower after communication has taken place.[15] Columns (4) and (5) in Table 3 report the results of running regression (2) using rank differences (based on average ratings) instead of rating differences as the outcome variable. There are 66 participants in treatment **OPENG**, so 1 is the best and 66 the worst possible rank. The table shows that women are ranked about three to four places lower ($\beta_2 + \beta_3$) and men about three to four places higher on average postcommunication. The difference is highly statistically significant.

Figure 2 (Panel b) shows the cumulative distribution function of differences in rank pre- and postcommunication. The figure shows that men tend to gain more than women in the sense of first-order stochastic dominance. More than 60% of men gain at least one position with one man gaining 25 positions postcommunication. By contrast, less than 25% of women gain in rank postcommunication. Online Appendix Table E.7 shows the top 10 rated summaries pre- and postcommunication. This table also shows that on average women lose three positions in the ranking postcommunication. Although the number of women placed in the top 5% remains the same postcommunication, the number of women in the top 10%, 20%, and 50% decreases.

4.3. *Heterogeneity.* We have a look at the heterogeneity of the effect. First, we ask whether it is the women with higher-rated or those with lower-rated surveys precommunication who lose out to men in terms of their ranking postcommunication. This is interesting for two reasons. First, implications for organizational design can differ depending on whether it is the "best" women who lose out or the "worst" women. In the context of hiring or promotion, where only the best candidates are selected, the former is probably of more concern than

[15] Precommunication women are rated about 0.12 points higher on average. Women's ratings stay the same postcommunication on average and the chance that a man's rating is increased by more than 0.12 postcommunication is about 40%.



NOTE: Error bars show standard errors of the mean.

FIGURE 3

HETEROGENEITY IN RANK DIFFERENCES DEPENDING ON PRECOMMUNICATION RATINGS AND GENDER
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

the latter. If the context is one where worst performers are penalized, then the latter should be of more concern. Second, understanding heterogeneity can address a potential concern for ceiling effects. To the extent that women's summaries are better rated precommunication than those written by men, there is less scope for their ratings to improve postcommunication. If we found that the effect comes predominantly from top rated women losing out, then ceiling effects could be partially responsible for this.

Figure 3 shows rank differences in treatments **OPEN** and **OPENG** as a function of precommunication ratings and gender. In treatment **OPEN** rank differences tend to be small across all levels of precommunication ratings and no different between men and women. In treatment **OPENG**, women lose in rank and men gain. In line with the idea of ceiling effects, women with very low precommunication ratings (≤ 4) do not lose many positions, whereas men with such ratings can gain up to seven positions on average. Gains and losses in rank are approximately stable for precommunication ratings between 5 and 10. Women lose between three and five positions on average and men gain zero to five positions across the distribution. Importantly, the gender difference in positions gained and lost seems approximately stable for all ratings between 2 and 10 and it seems to exist across all levels of precommunication ratings.

Are biases more severe if raters are male or female? Existing literature has been inconclusive on this question. Bagues and Esteve-Volart (2010) have shown that the bias against women is biggest if there are more women in the committee. On the other hand, Goldin and Rouse (2000) found gender biases in a setting where committees consisted of mostly men and Bagues et al. (2017) and Mengel et al. (2019) have found that men tend to display bigger biases in promotion decisions and performance evaluations, respectively. Online Appendix Table E.3 shows that in our setting gender biases exist for both rater genders without statistically significant differences. It should also be noted that the bias is in both cases (male and female raters), a bias against women. As such it is fundamentally different from in-group bias, where people treat in-group members more favorably and where hence we would expect men to favor men and women to favor women (Tajfel and Turner, 1986; Chen and Li, 2009).

4.4. *Mechanisms.* In this subsection, we present a series of analyses of the deliberation process to gain a better understanding of the source of gender bias with committee deliberation. Committee deliberation has a number of features that are not present in individual

TABLE 4
FINAL (POSTCOMMUNICATION) RATINGS IN THE **MIN** TREATMENTS REGRESSED ON AUTHOR GENDER, A DUMMY INDICATING WHETHER THE TREATMENT IS ONE WHERE GENDER IDENTITY IS REVEALED, AND THE INTERACTION OF THE LATTER WITH AUTHOR GENDER

	Minimal Deliberation		
	(1)	(2)	(3)
male (β_1)	0.229	−0.055	−0.152
	(0.186)	(0.187)	(0.209)
δ_G (β_2)	0.343	−0.702∗	−0.621
	(0.219)	(0.361)	(0.389)
$\delta_G \times$male (β_3)	−0.044	0.210	0.254
	(0.275)	(0.270)	(0.308)
Constant	6.158***	6.744***	6.817***
	(0.151)	(0.399)	(0.406)
Drop Constant Raters	–	–	✓
Session FE	–	✓	✓
Demographics	–	✓	✓
p-values			
Test $\beta_1 + \beta_3 = 0$	0.7774	0.5490	0.6160
Observations	1,161	1,161	1,116
R-squared	0.006	0.138	0.144

NOTES: Except for column (1) we also control for author demographics (age, whether the author is a student and nationality fixed effects) and session fixed effects. In column (3) we additionally drop all raters who enter the same rating for all summaries (see footnote). Robust standard errors clustered at the rater level.
Robust standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

decisions: (i) the possibility to deliberate, that is, persuade or convince others of one's position, (ii) the possibility to exchange privately held information, and (iii) incentives to reach an agreement. In our treatments **MIN**, **NO-IAG,** and **NO,** we sequentially shut down one of these features of a typical committee deliberation process. We discuss these treatments in Subsections 4.3.1. and 4.3.2. Afterward, we focus in some detail on the revision process (4.3.3), the chats (4.3.4), and committee composition effects (4.3.5).

4.4.1. *Minimal deliberation.* We first discuss our minimal-deliberation treatments, which are identical to the open-deliberation treatments except for the fact that there is no chat. Participants can however still see the ratings given by other committee members, which we refer to as "minimal communication."

Table 4 shows regression (1) using final (postcommunication) ratings in the **MIN** treatments as endogenous variable. The minimal treatments do not seem to induce a gender bias. The coefficient β_3 is small in absolute value ($\beta_3 \in [-0.04, 0.25]$) and is far from statistical significance in all specifications. Columns (1)–(3) in Online Appendix Table E.2 show the difference between pre- and postcommunication ratings. Columns (4) and (5) use rank as endogenous variable, where rank is based on the average rating received by a participant; 1 is the best rank and 66 (63) the worst rank in treatments **MIN** and **MING**, respectively. The table shows that there are also very few treatment differences in how men and women are ranked before and after minimal communication. We summarize that there is no evidence of gender bias in the minimal-deliberation treatments.

4.4.2. *NO communication.* What about the treatments where all communication is shut down? Table 5 shows the results of estimating Equation (1) for these treatments. Without incentives to agree, that is, with purely individual decisions (**NO** treatments) men gain around 0.4–0.6 rating points over women when gender identity is revealed (β_3). This difference is statistically significant in the specification without controls, but loses statistical significance once demographic controls and session fixed effects are included (column (2)) and constant raters

TABLE 5
BASELINE RATINGS REGRESSED ON AUTHOR GENDER, A DUMMY INDICATING WHETHER THE TREATMENT IS ONE WHERE GENDER IDENTITY IS REVEALED, AND THE INTERACTION OF THE LATTER WITH AUTHOR GENDER

	NO Treatments			NO-IAG Treatments		
	(1)	(2)	(3)	(4)	(5)	(6)
male (β_1)	−0.419**	−0.343*	−0.322*	0.330	0.198	0.259
	(0.174)	(0.176)	(0.173)	(0.204)	(0.219)	(0.225)
δ_G (β_2)	−0.073	−0.073	−0.084	−0.491	0.189	0.210
	(0.213)	(0.218)	(0.303)	(0.336)	(0.395)	(0.397)
$\delta_G \times$male (β_3)	0.638**	0.521*	0.444	−0.204	−0.158	−0.218
	(0.294)	(0.287)	(0.287)	(0.484)	(0.500)	(0.503)
Constant	6.107***	7.119***	7.622***	6.337***	8.273***	8.244***
	(0.125)	(0.546)	(0.615)	(0.163)	(0.785)	(0.787)
Drop Constant Raters	–	–	✓	–	–	✓
Session FE	–	✓	✓	–	✓	✓
Demographics	–	✓	✓	–	✓	✓
p-values						
Test $\beta_1 + \beta_3 = 0$	0.1130	0.1628	0.1270	0.7774	0.5497	0.4071
Observations	1,080	1,080	1,062	1,134	1,134	1,125
R-squared	0.008	0.043	0.035	0.019	0.087	0.088

NOTES: Except for columns (1) and (4) we also control for author demographics (age, whether the author is a student, and nationality fixed effects) and session fixed effects. In columns (3) and (6) we additionally drop all raters who enter the same rating for all summaries (see footnote). Robust standard errors clustered at the rater level.
Robust standard errors in parentheses.
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

are dropped (column (3)). With incentives to agree (**NO-IAG** treatments), gender bias becomes negative with men losing between 0.1 and 0.2 rating points over women when gender identity is revealed, but this difference is far from statistical significance.[16]

We conclude that there is no evidence of gender bias in the committee without deliberation but with incentives to agree (**NO-IAG** treatments). Also in the purely individual treatment (**NO**), there is no robust evidence of gender bias, though a moderate amount of gender bias cannot be ruled out. Shutting down communication hence seems to eliminate bias at least as long as incentives to agree are provided. For a variety of reasons shutting down communication might be neither desirable nor feasible, though. We discuss interventions that do not rely on shutting down communication in Section 5. Before we do so, however, we analyze the committee deliberation process in treatment **OPENG** in some more detail.

4.4.3. *Who revises postcommunication?*. We ask whether those who seem more "biased" change their opinion more or less compared to others postcommunication. In order to identify potentially biased participants, we compute the difference between rater j's and the committee's average ratings for male authors as well as for female authors. Our measure of individual gender bias "Bias$_j$" is the difference between the two. The higher "Bias$_j$" the more rater j favors men relative to the committee. If "Bias$_j$" is negative, then rater j favors women over men more than the committee on average.[17]

[16] One possible concern with this treatment could be that participants try to coordinate on salient outcomes (e.g., 10) in order to achieve agreement without the possibility to communicate. This does not seem to be the case. The variance of individual ratings is 3.74 in the **NO-IAG** condition compared to 3.48 in **NO**, suggesting that participants choose different ratings at least as often in **NO-IAG** compared to **NO**. It should also be noted that although in other treatments there are some constant raters (see footnote), there are none in the **NO-IAG** conditions.

[17] Online Appendix Figure F.4 shows the distribution of "Bias$_j$" across our six main treatments. The figure shows that individual biases are bigger when gender identity is known. This speaks against an argument that people are trying to "equalize" the ratings they give to men and women. When gender identity is known people tend to make bigger differences in the ratings they give to the two genders compared to when gender identity is not known.

TABLE 6
COEFFICIENTS FROM SEPARATE OLS REGRESSIONS OF ABSOLUTE PERCENTAGE CHANGE IN RATINGS POSTCOMMUNICATION REGRESSED ON DIFFERENT DEFINITIONS OF DUMMY FOR "EXTREME" OPINIONS

	Upward Revisions		Downward Revisions	
	MING	OPENG	MING	OPENG
"Bias$_j$"	0.005	−0.173**	−0.008	−0.004
p-value F-test	0.0461		0.9121	

NOTE: All regressions control for standard deviation of precommunication ratings within a group and include session fixed effects.

TABLE 7
CHAT PARTICIPATION IN TREATMENT **OPENG**. NUMBER OF INTERJECTIONS MADE BY WOMEN AND MEN; SHARE OF MALE INTERJECTIONS IN A CONVERSATION DEPENDING ON THE SHARE OF MALES IN A GROUP; PARTICIPATION RATE OF WOMEN AND MEN (PARTICIPATION = 1 IF A PARTICIPANT SAYS SOMETHING AT LEAST ONCE); THE AVERAGE LENGTH OF THE CONVERSATION IN TERMS OF NUMBER OF INTERJECTIONS AND THE SHARE OF TIMES THE LAST INTERJECTION CAME FROM A MAN

			Share Male Participants			
	Women	Men	0	$\frac{1}{3}$	$\frac{2}{3}$	1
Number of Interjections	140	131	–	–	–	–
Share Male Interjections	–	–	0	0.333	0.692	1
Participation Rate	100%	100%	–	–	–	–
Length Conversation	–	–	11.75	15	15.666	9
Last Word Male	–	–	0	0.2	0.888	1

We then ask whether those with more biased opinions, as defined above, revise more or less than others. Table 6 shows the coefficients of four different regressions where the endogenous variable is the amount by which a rating is changed postcommunication. Regressions are run separately for upward and downward revisions and for the two communication variations **MIN** and **OPEN**.

The table shows some interesting patterns. The more gender-biased participants are, the less likely they are to revise their rating upward in **OPENG**, but not in **MING**. This suggests that greater stickiness of biased opinions could be one of the channels through which gender bias persists under open deliberation. One implication of "sticky extreme opinions" is that men will particularly benefit in groups where the standard deviation of priors (initial ratings) is high. Online Appendix Table E.8 evaluates this hypothesis. In line with the stickiness hypothesis, men benefit particularly if the standard deviation in precommunication ratings is high.

4.4.4. *Chat analysis.* In this subsection, we analyze the chats from treatments **OPEN** and **OPENG**. Table 7 summarizes some key statistics regarding chat participation. The table shows that all participants say something at least once in every chat (participation rate is 100%) and men and women speak up about equally often. There are also no statistically significant differences in the length of conversations or in who "has the last word."

We now ask whether those with potentially biased opinions tend to take leadership in conversations. More specifically, we ask whether those with more biased opinions (according to "Bias$_j$") are more likely to speak first in a chat and whether they tend to speak more often. Table 8 shows the results. In **OPEN,** we do not find statistically significant effects. In **OPENG,** we find that raters with more "biased" opinion are more likely to speak first (column (2), $p = 0.1080$). However, the effect is just outside of conventional levels of statistical significance.[18]

[18] We also ask whether those who speak first have more influence on the outcome, specifically on the postcommunication average rating in a group. Table E.9 in Online Appendix E shows regressions where this outcome is regressed on precommunication ratings as well as interactions with a dummy indicating that a participant spoke first or

TABLE 8
DUMMY INDICATING THAT PARTICIPANT SPEAKS FIRST (COLUMNS (1) AND (2)) AS WELL AS SHARE OF INTERJECTIONS DURING A CHAT (COLUMNS (3) AND (4)) REGRESSED ON MEASURES "EXTREME$_j$" AND "BIAS$_j$" IN SEPARATE REGRESSIONS

	Speak First		Share of Interjections	
	OPEN	OPENG	OPEN	OPENG
	(1)	(2)	(3)	(4)
"Bias$_j$"	−0.017	0.042	−0.004	−0.002
	(0.042)	(0.011)	(0.012)	(0.006)

NOTES: The table shows only the coefficient on the measures of bias.
Robust standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We also analyzed what was said during chats. Most chats are on topic, there is no discriminatory language used, and not a single case where one of the raters openly argues with the gender of the person who wrote the summary. We conducted a sentiment analysis (Turney, 2002; Pang et al., 2002) to understand whether word analysis of what was communicated in **OPENG** is consistent with the fact that mostly mens' ratings tend to improve after open communication.[19] Online Appendix Table E.10 shows that the distribution of positive, neutral, and negative sentiment is approximately balanced for all gender combinations. There are more negative sentiments expressed, though, when the summary rated was written by a female participant compared to when it was written by a man. The sentiment analysis hence is consistent with the finding that open communication is more likely to lead to upward revisions for men than for women.

4.4.5. *Committee composition and gender effects within the committee.* This subsection focuses on the gender composition of the committee. Existing literature has been inconclusive on the role of committee composition for gender bias, with some finding no correlation (Vernos, 2013; Auspurg et al., 2017), some that more men on the committee is worse for female candidates (De Paola and Scoppa, 2015), and some that more women on the committee is harmful to female candidates (Bagues and Esteve-Volart, 2010; Deschamps, 2020).[20]

To start with, we ask how gender bias (β_3) differs across male-dominated and female-dominated groups. Table 9 shows that there is no statistically significant difference between gender bias in female-dominated and male-dominated groups.

We also ask whether participants show differential behavior in precommunication ratings and if they revise differentially depending on the committee composition. Online Appendix Table E.14 shows average precommunication ratings for different committee compositions. The table shows that neither female nor male raters display consistently different precommunication ratings for different committee compositions. The only statistically significant effect is that men give somewhat higher ratings to male authors in all-male committees. We also do

that a participant contributed more than half of the conversation. The regressions show some positive effect of both "speaking first" and "share of interjections" on the outcome, though the effect is not statistically significant ($p = 0.139$ and $p = 0.140$, respectively).

[19] Sentiment analysis refers to the use of machine learning algorithms to identify and extract subjective information in source materials. We classify our chat data according to whether they have a positive, neutral, or negative semantic orientation. A phrase has a positive semantic orientation when it has good associations (e.g., "cool" or "sweet" in our chats) and a negative semantic orientation when it has bad associations (e.g., "sigh" or "guys, it's a bad summary"). Relatively informal chats such as ours can present a challenge for sentiment analysis, though techniques have improved in recent years to deal with such conversations (Thelwall et al., 2010) and have been shown to be effective when applied, for example, to analyze reviews of movies or restaurants.

[20] Interestingly Deschamps (2020) finds that the negative effect on female candidates is concentrated in committees chaired by men.

TABLE 9
GENDER BIAS (β_3 IN EQUATION (1)) DEPENDING ON COMMITTEE COMPOSITION

	MIN Treatments		OPEN Treatments	
Male Committee Members	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
β_3	−0.108	0.316	0.649*	0.863*
p-value F-test	0.3268		0.6526	
N	558	477	630	450

NOTE: Committees with zero or three male committee members are too infrequent and hence omitted.

not find substantial differences in terms of when raters adjust their ratings more (Online Appendix Table E.15).

5. INTERVENTIONS

In this section, we analyze evidence from two interventions we conducted in the lab to test whether they can reduce gender bias under open deliberation. Under the first intervention, discussed in Subsection 5.1, we randomize the order of speaking. Under the second intervention, discussed in Subsection 5.2, we inform participants about gender bias in previous sessions.

5.1. *"Order of Speaking" Intervention.* Our first intervention targets the order of speaking. Economists have long recognized that the order of speaking in committee deliberation is not irrelevant and hence it is one dimension that can be targeted to affect committee decisions; see, for example, Ottaviani and Soerensen (2001) or Fershtman and Segal (2020). As we have also seen possible evidence ($p = 0.1080$) that more biased committee members tend to speak first, targeting the order of speaking seems a worthwhile intervention. In treatment **OPEN-RAND**, we randomly picked one of the three group members to speak (send a chat message) first. The other two group members were not able to send messages until this firstrating received in the message had been sent. This section studies whether this intervention is able to reduce gender bias.

Table 11 shows results from regression (1) run on the final ratings from the **OPEN-RAND** treatments. The table shows that gender bias is substantial with coefficient $\beta_3 > 1$ in all specifications. This is the biggest gender bias identified across all our treatments. An F-test comparing the size of gender bias in column (3) with the corresponding specification for treatment **OPEN** has a p-value of $p = 0.0477$. Panel a in Figure 4 shows the distribution of standardized ratings and illustrates the extent of gender bias across the distribution under this intervention.[21]

The evidence from this subsection shows that the intervention was not successful. Randomizing who speaks first is not sufficient to reduce gender bias. One possible reason for this is that the strength of statements is important. If those randomly selected first speakers, who would not endogenously speak first, provide weak statements then the effect of randomizing the order of speaking could be easily washed out. In fact it might even be possible that gender bias gets worse in these cases, as those with biased views see more opportunities to swing the committee opinion in case the first statement is weak. Figure 5 provides suggestive evidence in this regard. The figure shows pre- and postcommunication bias over time, specifically for the first three, middle three, and last three ratings a committee conducts. What can be seen is that in **OPEN-RAND**, unlike in any other open treatment, gender bias is strongly increasing over time. This could suggest that gender-biased committee members learn to exploit first

[21] Online Appendix Table E.11 shows some of the heterogeneity underlying the gender bias in this treatment. Specifically, the table shows that gender bias is at least as big if the first speaker is female compared to when they are male. The table also shows that, as in other treatments, gender bias is present for both female and male raters.

TABLE 10
FINAL RATINGS REGRESSED ON AUTHOR GENDER, A DUMMY INDICATING WHETHER THE TREATMENT IS ONE WHERE GENDER IDENTITY IS REVEALED, AND THE INTERACTION OF THE LATTER WITH AUTHOR GENDER

	OPEN-RAND Treatments		
	(1)	(2)	(3)
male (β_1)	−0.010	−0.352*	−0.354*
	(0.202)	(0.193)	(0.193)
δ_G (β_2)	−0.830***	1.097***	1.084***
	(0.296)	(0.360)	(0.361)
$\delta_G \times$ male (β_3)	1.154***	1.420***	1.438***
	(0.369)	(0.352)	(0.356)
Constant	5.889***	7.415***	7.428***
	(0.164)	(0.382)	(0.385)
Drop Constant Raters	–	–	✓
Session FE	–	✓	✓
Demographics	–	✓	✓
p-values			
Test $\beta_1 + \beta_3 = 0$	0.0003	0.0003	0.0003
Observations	1,080	1,080	1,071
R-squared	0.024	0.238	0.239

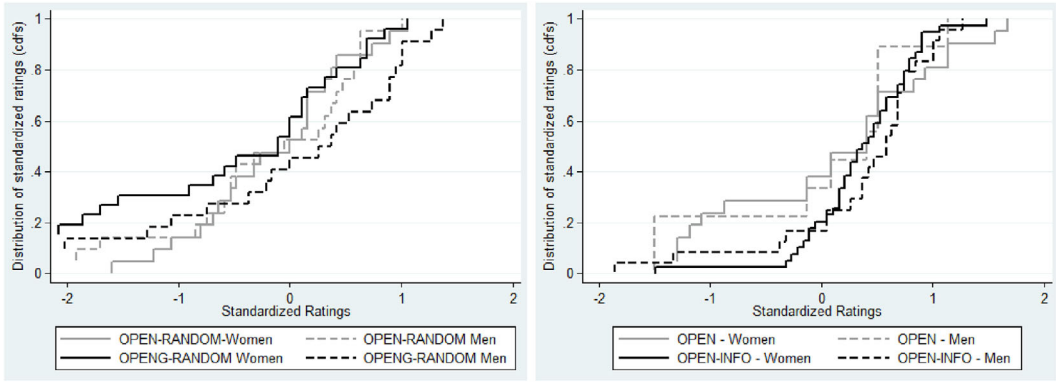
NOTES: Except for column (1) we also control for author demographics (age, whether the author is a student, and a nationality indicator) and session fixed effects. Column (3) drops one constant rater. Standard errors clustered at the rater level.
Robust standard errors in parentheses.
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

TABLE 11
FINAL RATINGS REGRESSED ON AUTHOR GENDER, A DUMMY INDICATING WHETHER THE TREATMENT IS ONE WHERE GENDER IDENTITY IS REVEALED, AND THE INTERACTION OF THE LATTER WITH AUTHOR GENDER

	OPENG-INFO		
	(1)	(2)	(3)
male (β_1)	−0.151	−0.151	−0.333*
	(0.0931)	(0.141)	(0.141)
δ_G (β_2)	0.302	0.861***	0.838***
	(0.191)	(0.284)	(0.285)
$\delta_G \times$ male (β_3)	0.091	0.087	0.207
	(0.218)	(0.228)	(0.228)
Constant	6.225***	7.006***	6.925***
	(0.0802)	(0.580)	(0.597)
Drop Constant Raters	–	–	✓
Session FE	–	✓	✓
Demographics	–	✓	✓
p-values			
Test $\beta_1 + \beta_3 = 0$	0.7645	0.5202	0.5540
Observations	1,107	1,107	1,080
R-squared	0.008	0.088	0.093

NOTES: Except for column (1) we also control for author demographics (age, whether the author is a student, and a nationality indicator) and session fixed effects. Standard errors clustered at the rater level.
Robust standard errors in parentheses.
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

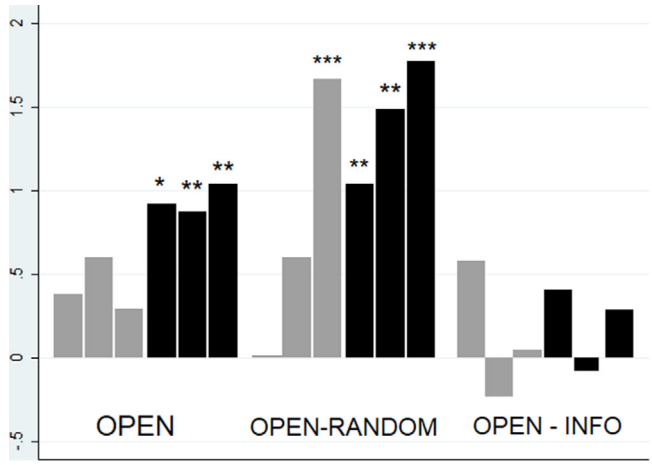
statements to move the committee opinion more toward their own. In order to explore this possibility further, we also assembled some direct evidence on "weak" statements. After explaining some context, we asked 117 participants in an online survey to rate the first chat messages according to whether they thought they constituted a "strong" or a "weak" statement. If we then look at gender bias in groups where first messages are rated as "weak" by fewer than



(a) OPEN-RAND (b) OPEN-INFO

FIGURE 4

DISTRIBUTION OF STANDARDIZED RATINGS IN OPEN TREATMENTS WITH RANDOMIZED ORDER OF SPEAKING (PANEL A) AND IN THE TREATMENTS WITH INFORMATION INTERVENTION (PANEL B) [COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]



NOTES: Three bars in each category show the estimated coefficient for the first, middle, and last three ratings conducted (from left to right). Specifications where demographics are controlled for and constant raters dropped.

FIGURE 5

THE FIGURE SHOWS THE SIZE AND STATISTICAL SIGNIFICANCE OF THE ESTIMATED COEFFICIENT β_3 FOR PRECOMMUNICATION RATINGS (GRAY BARS) AND POSTCOMMUNICATION TREATMENTS (BLACK BARS) FOR THE THREE TREATMENT CONDITIONS, **OPEN**, **OPEN-RAND,** AND **OPEN-INFO**
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

half or more than half of survey participants, we do find that gender bias is stronger in the latter group (Online Appendix Table E.12). The difference is not statistically significant, though.

Overall, these results show that detail can matter when trying to regulate the order of speaking in committees. Care needs to be applied in practice when using these types of interventions to avoid that they backfire and produce possibly worse biases than in the absence of an intervention.

5.2. *Information Intervention.* Our second intervention is an information intervention, where participants are made aware of gender bias in previous sessions prior to entering their ratings. There is a belief in management literature that bias awareness reduces biases and

some evidence that this could indeed be the case (Beshears and Gino, 2015). Most of the examples in the literature, though, concern non–committee decision making. Boring and Philippe (2017), for example, found that providing students with information that past teaching evaluations had been gender biased reduced the bias particularly among male students. Pope et al. (2018) find that awareness of racial bias among professional basketball referees eliminates the bias. Similarly Devine et al. (2012) found that attending a 12-week course to raise awareness of implicit bias did reduce implicit racial bias among non-black undergraduate students. Short online training, however, only showed limited effects in Chang et al. (2019). Burnell et al. (2018) find that awareness of bias is not enough to de-bias teaching evaluations in their study, though they do not use an explicit information intervention. Also Kalev et al. (2006) do not find diversity training to be very effective in increasing the share of women or black employees in management at U.S. firms.[22] We did not find prior evidence on the effectiveness of an information intervention in committee decision making. Studying such an intervention is all the more interesting, though, as committees are sometimes made aware of gender bias in practice, as in some grant panels in the Netherlands following a study about potential gender bias in research grant awards (Van der Lee and Ellemers, 2015).

In order to test the effectiveness of such interventions, we conduct treatments that are identical to the **OPENG** treatments, but where participants can see the following statement on the screen where they enter their precommunication ratings: "*In previous sessions of this study there has been evidence that women receive too low ratings. Please be aware of gender bias when entering your rating*."[23,24] We conducted this treatment only in the **G**-variation, as it seems to make little sense to do so in a gender-blind treatment. As a gender-blind condition we hence use the data from treatment **OPEN**. Table 11 shows the results of running regression 1 on these treatments. The table shows that there is no gender bias in this treatment with the coefficient β_3 ranging between 0.09 and 0.20, none of which is statistically significant. Panel b in Figure 4 shows the distribution of standardized ratings in this treatment and illustrates that there is no gender bias.[25]

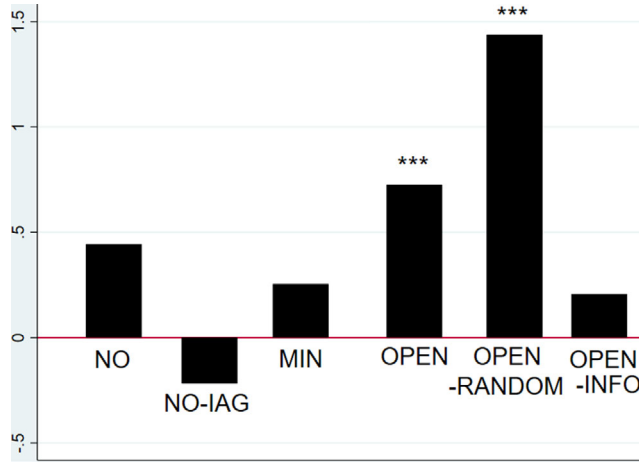
One concern with these types of interventions could be that they have "too strong" an effect and end up harming the previously favored side, in this case men. At least in our case, this does not seem to be the case with relatively small coefficients β_3. Note also that there could be experimenter demand effects in this treatment. This is also the case in the applications mentioned above. Informing students about bias in prior teaching evaluations (as in Boring and Philippe, 2017) or informing grant panel members about bias in decisions by previous committees (as in some grant panels following Van der Lee and Ellemers, 2015) very likely conveys a "demand" or an expectation set by the university administering teaching evaluations or by the funding body giving out research grants. Hence these types of demand effects here are very much part of the story. It is an empirically open question how people react to these "demands." Some might comply with the "demand," whereas others might resent that the "demand" is being made and might even lean to the opposite direction as a result. Hence, another concern with these types of interventions is that they could lead to polarization, with some raters complying and giving ratings that are favoring women, whereas others may react

[22] A different question, that we cannot address here, is how long-lasting these effects are. Macrae et al. (1994), for example, find that stereotypes are not reduced in the long term after these type of interventions.

[23] Note that this statement combines information ("...there has been evidence...") with an appeal ("... please be aware..."). As any statement informing about gender bias will implicitly contain an appeal (because it is implicitly understood that "gender bias" is "bad"), we decided to include the second component explicitly.

[24] One concern with this statement could be that it makes participants aware that the purpose of the study is to investigate gender bias. Although we cannot rule out this possibility, we can at least report that, as in the other treatments, participants did not mention gender in their answers to an open-ended question at the end of the experiment asking what they believe the study was about.

[25] If we rerun the regression reported in Table 11 using precommunication ratings as endogenous variable we find values for β_3 of 0.069 (without controls) and −0.111 (with controls). Hence providing participants with information does not lead to gender bias precommunication and, more importantly, also ensures that it does not appear postcommunication either.



NOTE: Coefficients from specifications with demographic controls, session fixed effects and where constant raters are dropped.

FIGURE 6

ESTIMATED GENDER BIAS COEFFICIENT (β_3) ACROSS ALL TREATMENTS
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

adversely to a message that might be perceived as patronizing and provide ratings that favor men. We do not find clear evidence of such polarization. The consensus rate (share of rounds where all three raters agree) is 0.39 in **OPENG-INFO** as opposed to 0.37 in **OPENG** and the standard deviation of ratings is also similar across these two variations with 0.86 in **OPENG** as opposed to 1.10 in **OPENG-INFO**. Although the higher standard deviation in **OPENG-INFO** is in line with polarization, the two standard deviations are not statistically different.

Finally, Figure 5 shows interesting differences in dynamics across the three different open communication variations. The figure shows the size and statistical significance of the estimated coefficient β_3 for precommunication ratings (gray bars) and postcommunication treatments (black bars) for the three treatment conditions, **OPEN**, **OPEN-RAND**, and **OPEN-INFO**. From left to right the three bars in each category show the estimated coefficient for the first, middle, and last three ratings conducted. As discussed above, in the **OPEN** treatments we do not see very clear dynamic patterns. Gender bias is higher postcommunication than precommunication, but there is no clear trend across the nine rounds of ratings. This is very different in the **OPEN-RAND** condition, where gender bias is strongly increasing over time. This is in line with the intuition developed above where biased participants learn to exploit weaker initial statements over time. In the **OPEN-INFO** condition, we do see some evidence of gender bias in the first three rounds prior to communication (albeit not statistically significant), which disappears postcommunication and in all subsequent rounds.

6. CONCLUSIONS

We used a combination of lab experiments and online surveys to study how deliberation contributes to gender biases. If—in line with the literature showing groups to learn faster and make more rational strategic decisions (Cooper and Kagel, 2005; Kocher and Sutter, 2005)—committee deliberation leads to "better" decisions also in terms of reducing bias, then we should see values of β_3 closer to zero with committee deliberation. If, by contrast, deliberation leads to groupthink, increased extremism, or polarization (Baldassari and Bearman, 2007; Cason and Mui, 1997), then it is possible that β_3 increases with committee deliberation.

Figure 6 summarizes our results. There is no statistically significant gender bias present in the treatments without communication. Incentives to agree if at all decrease gender bias. This

is true without as well as with minimal communication. Under open communication, by contrast, gender bias is strong and highly statistically significant. In the latter case, 60% of ratings received by men are revised upward after communication compared to only 25% of ratings received by women. We also tested two interventions. Randomizing the order of speaking did not reduce gender bias, but an information intervention, where participants are informed of gender bias in prior sessions, was successful in eliminating bias.

There are a number of possible implications of our findings for committee decision making that could be explored in future research. One finding from our study is that institutional detail can matter. Although under minimal communication there is no gender bias, it is substantial under open communication. Some committees collect individual ratings or opinions of all committee members *before* a meeting and use them to prerank candidates. Whether such interventions will mitigate gender biases is unclear and seems an interesting avenue for future research. Bohnet et al. (2016) have recently found that individual biases are reduced when people are evaluated jointly instead of separately for promotion or bonuses. As many of these decisions are made in committees where deliberation is possible (and in many cases unavoidable), whether or not one would recommend joint evaluation will also depend on whether biases are smaller under joint evaluation also in the presence of deliberation. This seems another interesting question for future research.

When interpreting our results it should be noted that they are obtained in a setting where the stakes for committee members to rate candidates correctly are relatively low. As such, the more relevant applications are cases where the marginal decision involves a choice between candidates of similar quality and where, hence, mistakes are not too costly for the committee.[26] It is an open question whether we would see the same patterns of gender bias in cases where the cost of a mistake can be very high.

It should also be kept in mind that, although we have found that unstructured open communication seems conducive to gender bias, this form of deliberation could have other advantages. Decision makers will have both the quality and speed of decisions in mind when designing committee deliberation rules, as well as possibly the level of consensus or agreement required to reach a decision. Open deliberation could, for example, lead to higher quality of ratings overall, irrespective of author gender.[27] We also found that the rate of agreement is higher under open communication compared to minimal communication, which can be an important consideration for committee decisions that require a certain level of consensus. Such possible other advantages need to be kept in mind when making policy recommendations.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure C.1: Age Distribution of participating men and women.
Table C.1: Balancing tests by treatment.
Table C.2: Balancing tests by whether treatment gender blind.
Figure D.1: Screenshot from Survey I.
Table D.1: Pairwise correlation between (pre-communication) rating received in the experiment and rating received in survey.

[26] Note that although mistakes are not too costly for the committee, they can be very costly for the candidate who ends up not being hired, as the analysis in Subsection 6.2 has demonstrated.

[27] It is hard to assess whether open communication indeed yields a better quality of ratings as we do not have an objective measure of quality. If we compare the final ratings with participants' own assessments of their summaries (elicited at the end of the experiment), we find that final ratings are much closer to own assessments with open communication compared to the no-communication condition, and slightly closer compared to minimal communication. However, the direction of causality of this effect is unclear.

Figure D.2: Frequency of answers to the question whether women or men performed better in the task.
Figure D.3: Frequency of answers to the question whether younger or older participants performed better in the task.
Table E.1: Main regression pooling all deliberation variations (non intervention treatments only and without NO-IAG).
Table E.2: OLS Regression of difference between post and pre- communication ratings (columns (1)-(3)) and rank based on average ratings (columns (4)-(5)) on author-gender, a dummy indicating whether the treatment is one where gender-identity is revealed and the interaction of the latter with author gender.
Table E.3: Main regression separately for different rater genders.
Table E.4: Final ratings regressed on author-gender, a dummy indicating whether the treatment is one where gender-identity is revealed and the interaction of the latter with author gender.
Table E.5: Pre-communication ratings regressed on author-gender, a dummy indicating whether the treatment is one where gender-identity is revealed and the interaction of the latter with author gender.
Table E.6: Regression on pre and post communication ratings differences separately for different rater genders.
Table E.7: Top 10 ranked authors (performance and gender) pre-communication and post-communication in treatment OPENG as well as number of women in top 5%; 10% etc.
Table E.8: Difference between pre- and post communication ratings (δ_i = RRi-Ri) regressed on gender of person rated. Sample split depending on standard deviation in pre-communication ratings.
Table E.9: Average post-communication ratings regressed on pre-communication ratings interacted with a dummy indicating whether rater spoke first (columns (1) and (2)) or with dummy indicating whether rater dominated conversation (whether more than half of interjections came from a rater) - columns (3) and (4).
Table E.10: Share of messages classified as positive, neutral or negative by the machine learning algorithm depending on the gender (M for male, F for female) of the rater and the author.
Table E.11: Final ratings regressed on author-gender, a dummy indicating whether the treatment is one where gender-identity is revealed and the interaction of the latter with author gender.
Table E.12: Final ratings regressed on author-gender, a dummy indicating whether the treatment is one where gender-identity is revealed and the interaction of the latter with author gender.
Table E.13: Final ratings regressed on author-gender, a dummy indicating whether the treatment is one where gender-identity is revealed and the interaction of the latter with author gender.
Table E.14: Average pre-communication ratings given by female and male raters depending on committee composition (number of male committee members 0-3) and author gender (male or female).
Table E.15: Average amounts by which female and male raters revise their ratings depending on committee composition (number of male committee members 0-3) and author gender (male or female).
Figure F.1: Gender bias with communication: distribution of standardized ratings in MIN and MING for both men and women.
Figure F.2: Average ratings for each avatar (black diamonds) as well as distribution of mean ratings across all participants playing with this avatar (hollow diamonds).
Figure F.3: Evolution of ratings (panel(a)), the standard deviation of ratings (panel (b)) and agreement (panel (c)) over time within groups of three raters.
Figure F.4: CDF of individual bias measure across the six main treatments.

REFERENCES

Altonji, J., and R. Blank. "Race and Gender in the Labor Market," in Ashenfelter, O. and D. Card (Ed.), *Handbook of Labor Economics*, Volume 3, Chapter 48 (Amsterdam: Elsevier Science B.V., 1999), 3143–99.

Ambrus, A., B. Greiner, and P. Pathak, "How Individual Preferences Are Aggregated in Groups: An Experimental Study," *Journal of Public Economics* 129 (2015), 1–13.

Anderson, L., and Charles A. Holt, "Information Cascades in the Laboratory," *American Economic Review* 87(5) (1997), 847–62.

Asch, S. E. "Opinions and Social Pressure," in E. Aronson (Ed.), *Readings about the Social Animal*. 13 (New York:Worth Publishers, 1995), 15–17.

Auspurg, K., T. Hinz, and A. Schneck, "Appointment Procedures as Tournaments: Gender-Specific Chances of Being Appointed as Professors," *Zeitschrift fuer Soziologie* 46(4) (2017), 283–302.

Bagues, M. F., and B. Esteve-Volart, "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment," *The Review of Economic Studies* 77(4) (2010), 1301–28.

———, M. S. Labini, and N. Zinovyeva, "Does the Gender Composition of Scientific Committees Matter?" *American Economic Review* 107(4)(5537) (2017), 1207–1238.

Baldassari, D., and P. Bearman, "Dynamics of Political Polarization," *American Sociological Review* 72 (2007), 784–811.

Benabou, R., "Groupthink: Collective Delusions in Organizations and Markets," *Review of Economic Studies* 80 (2013), 429–462.

Beshears, J., and F. Gino., "Leaders as Decision Architects," *Harvard Business Review* (2015, May). 17, 1–10.

Black, D., A. Haviland, S. Sanders, and L. J. Taylor, "Gender Wage Disparities Among the Highly Educated," *Journal of Human Resources* 43(3) (2008), 630–59.

Bohnet, I., A. van Geen, and M. Bazerman, "When Performance Trumps Gender Bias: Joint Versus Separate Evaluation," *Management Science* 62(5) (2016), 1225–34.

Booth, A., M. Francesconi, and J. Frank, "A Sticky Floors Model of Promotion, Gender and Pay," *European Economic Review* 47(2) (2003), 295–322.

Boring, A., "Gender Biases in Student Evaluations of Teachers," *Journal of Public Economics* 145 (2017), 27–41.

———, and A. Philippe, "Reducing Discrimination Through Norms or Information: Evidence from a Field Experiment on Student Evaluations of Teaching." TSE Working Paper, 2017.

Brandts, J., A. Giritligil, and R. Weber, "An Experimental Study of Persuasion Bias and Social Influence in Networks," *European Economic Review* 80 (2015), 214–29.

Burnell, N., I. Cojuharenco, and Z. Murad. "Now You See It Now You Don't: The Effect of Teaching Style and Seniority on Gender Bias in Teaching Evaluations," Mimeo, University of Surrey 2018.

Casari, M., and J. Zhang, "How Groups Reach Agreement in Risky Choices: An Experiment," *Economic Inquiry* 50(2) (2012), 502–15.

Cason, T., and V. Mui, "A Laboratory Study of Group Polarisation in the Team Dictator Game," *Economic Journal* 107 (1997), 1465–83.

Ceci, S. J., D. K. Ginther, S. Kahn, and W. M. Williams, "Women in Academic Science: A Changing Landscape," *Psychological Science in the Public Interest* 15(3) (2014), 75–141.

Chandrasekhar, A., H. Larreguy, and J. Xandri, "Testing Models of Social Learning on Networks: Evidence from a Lab Experiment in the Field," NBER Working Paper 21468, 2015.

Chang, E., K. Milkman, D. Gromet, R. Rebele, C. Massey, A. Duckworth, and A. Grant, "The Mixed Effects of Online Diversity Training," *PNAS* 116(16) (2019), 7778–83.

Chen, Y., and S. X. Li, "Group Identity and Social Preferences," *American Economic Review* 99(1) (2009), 431–57.

Coffman, K. B., C. Flikkema, and O. Shurchkov, "Gender Stereotypes in Deliberation and Team Decisions," Working Paper, Harvard University. 2020.

Cooper, D., and J. Kagel, "Are Two Heads Better Than One? Team Vs Individual Play in Signaling Games," *The American Economic Review* 95(3) (2005), 477–509.

———, and ———, "A Failure to Communicate: An Experimental Investigation of the Effects of Advice on Strategic Play," *European Economic Review* 82 (2016), 24–45.

Corrazini, L., F. Petrovesi, B. Petrovich, and L. Stanca, "Influential Listeners: An Experiment on Persuasion Bias in Social Networks," *European Economic Review* 56(6) (2012), 1276–88.

De Paola, M., and V. Scoppa, "Gender Discrimination and Evaluators' Gender: Evidence from the Italian Academia," *Economica* 82(325) (2015), 162–88.

Della Vigna, S., and D. Pope, "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy*, 126 (2018), 2410–2456.

Deschamps, P., "Gender Quotas in Hiring Committees: A Boon or a Bane for Women?" Mimeo, Sciences Po. 2020.

Devine, P., P. Forscher, A. Austin, and W. Cox, "Long-term Reduction in Implicit Racial Bias: A Prejudice Habit-Breaking Intervention," *Journal of Experimental Social Psychology* 48(6) (2012), 47–53
Fershtman, C., and U. Segal, "Social Influence in Committee Deliberations," Mimeo BU, 2020.
Ginther, D., and S. Khan, "Women in Economics: Moving Up or Falling Off the Academic Career Ladder? *Journal of Economic Perspectives* 18(3) (2004), 193–214.
Goeree, J., and L. Yariv, "An Experimental Study of Collective Deliberation," *Econometrica* 79(3) (2011), 893–923.
Goldin, C., and C. Rouse, "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians," *American Economic Review* 90(4) (2000), 715–41.
Grimm, V., and F. Mengel, "Experiments on Belief Formation in Networks," *Journal of the European Economic Association* 18(1) (2020), 49–82.
Holmes, J., and M. Meyerhoff, eds. *The Handbook of Language and Gender*. (Oxford: Wiley Blackwell Publishing, 2003).
Isenberg, D. "Group Polarization: A Critical Review and Meta-Analysis," *Journal of Personality and Social Psychology* 1141 (1986), 1141.
Janis, I. *Victims of Groupthink*. (Boston, MA: Houghton Mifflin, 1972).
Kalev, A., F. Dobbin, and E. Kelly, "Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies," *American Sociological Review* 71(4) (2006), 589–617.
Kocher, M., and M. Sutter, "The Decision Maker Matters: Individual Versus Group Behaviour in Experimental Beauty-Contest Games," *Economic Journal* 115(500) (2005), 200–23.
Krawczyk, M., and M. Smyk, "Author's Gender Affects Rating of Academic Articles: Evidence from an Incentivized, Deception-Free Experiment," *European Economic Review (Special Issue on Social Identity)* 90 (2016), 326–36.
Luhan, W., M. Kocher, and M. Sutter, "Group Polarization in the Team Dictator Game Reconsidered," *Experimental Economics* 12(1) (2009), 26–41.
MacNell, L., A. Driscoll, and A. N. Hunt, "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching," *Innovative Higher Education* 40(4) (2015), 291–303.
Macrae, C. N., G. Bodenhausen, A. Milne, and J. Jetten, "Out of Mind But Back in Sight: Stereotypes on the Rebound," *Journal of Personality and Social Psychology* 67(5) (1994), 808–17.
Malmstrom, M., A. Voitkane, J. Johansson, and J. Wincent, "Vc Stereotype About Men and Women Aren'T Supported by Performance Data," *Harvard Business Review* 22 (2018, March), 1–8.
Marder, N. "Gender Dynamics and Jury Deliberation," *The Yale Law Journal* 96(3) (1987), 593–612.
Masclet, D., N. Colombier, L. Denant-Boemont, and Y. Loheac, "Group and Individual Risk Preferences: A Lottery Choice Experimnent with Self-Employed and Salaried Workers," *Journal of Economic Behavior and Organization* 70 (2009), 470–84.
Mengel, F. "Gender Differences in Networking," *Economic Journal* 130(630) (2020), 1842–73.
———, J. Sauermann, and U. Zoelitz, "Gender Bias in Teaching Evaluations," *Journal of the European Economic Association* 17(2) (2019), 535–56.
Neumark, D., R. Bank, and K. D. V. Nort, "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics* 111(3) (1996), 915–41.
Ottaviani, M., and P. Soerensen, "Information Aggregation in Debate: Who Should Speak First?" *Journal of Public Economics* 81 (2001), 393–421.
Ozdemir, D., "Absence of Female Leaders: Do Group Dynamics Play A Role?" Mimeo, University of Essex 2018.
Pang, B., L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79–86, 2002.
Plante, I., M. Theoret, and O. Favreau, "Student Gender Stereotypes: Contrasting the Perceived Maleness and Femaleness of Mathematics and Language," *Educational Psychology* 29(4) (2009), 385–405.
Pope, D., J. Price, and J. Wolfers, "Awareness Reduces Racial Bias," *Management Science* 64(11) (2018), 4967–5460.
Rawls, J., *A Theory of Justice* (Cambridge, MA: Belknap, 1971), 128(613), 2131–2159.
Sandberg, A., "Competing Identities: A Field Study of In-Group Bias Among Prprofession Evaluators," *Economic Journal* in press (2018).
Sandberg, S. *Lean In: Women, Work and the Will to Lead*. (New York, NY: Knopf Publishers, 2013).
Schwarz, M., and L. Vesterlund, "Gender and Group Decision Making: Experimentally Connecting Individual and Group Beliefs," Mimeo, University of Pittsburgh 2020.
Shupp, R., and A. Williams, "Risk Preference Differentials of Small Groups and Individuals," *Economic Journal* 118 (2008), 258–83.
Spencer, S., C. Steele, and D. Quinn, "Stereotype Threat and Women's Math Performance," *Journal of Experimental Social Psychology* 35(1) (1999), 4–28.

Steele, C., S. Soencer, and J. Aronson, "Contending with Group Image: The Psychology of Stereotype and Social Identoty Threat," *Advances in Experimental Social Psychology* 34 (2002), 379–440.
Sunstein, C. R. "Deliberative Trouble? Why Groups Go to Extremes," *The Yale Law Journal* 110 (2000), 71–119.
Surowiecki, J., *The Wisdom of Crowds* (New York, NY: Anchor Books, 2004).
Swim, J., E. Borgida, G. Maruyama, and D. Myers, "Joan Mckay Versus Johnmckay: Do Gender Stereotypes Bias Evaluations?" *Psychological Bulletin* 105(3) (1989), 409–29.
Tajfel, H., and J. C. Turner.,*The Social Identity Theory of Inter-Group Behavior* (Chicago, IL: Nelson Hall, 1986).
Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment Detection in Short Informal Text," *Journal of the American Society for Information Science and Technology* 61(12) (2010), 2544–58.
Turner, M. E., and A. Pratkanis, "Twenty-Five Years of Group Theory and Research: Lessons from the Evaluation of a Theory," *Organizational Behavior and Human Decision Processes* 73 (1998), 105–15.
Turney, P., "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," in *Proceedings of the Association for Computational Linguistics*, 2002, 417–24.
Van der Lee, R., and N. Ellemers, "Gender Contributes to Personal Research Funding Success in the Netherlands," *Proceedings of the National Academy of Sciences of the United States of America* 112(40) (2015), 12349–53.
Vernos, I., "Quotas are Questionable," *Nature* 495(39) (2013), 1–12.