HARVARD | BUSINESS | SCHOOL

Gender Stereotypes in Deliberation and Team Decisions

Katherine Coffman
Clio Bryant Flikkema
Olga Shurchkov

Working Paper 19-069

Gender Stereotypes in Deliberation and Team Decisions

Katherine Coffman
Harvard Business School

Clio Bryant Flikkema

Olga Shurchkov
Wellesley College

Working Paper 19-069

Funding for this research was provided in part by Harvard Business School and Wellesley College Faculty Awards.

GENDER STEREOTYPES IN DELIBERATION AND TEAM DECISIONS[1]

Katherine Coffman[2] Clio Bryant Flikkema[3] Olga Shurchkov[4]

Harvard Business School Wellesley College

First Draft: November 2018

Current Draft: January 2021

Abstract: We explore how groups deliberate and decide on ideas in an experiment with communication. We find that gender biases play a significant role in which group members are chosen to answer on behalf of the group. Conditional on the quality of their ideas, individuals are less likely to be selected in gender incongruent domains (i.e. male-typed domains for women; female-typed domains for men). Individuals are also less likely to promote themselves when they are in the gender minority within their group. These patterns are not well-explained by objective or subjective differences in conversational behavior, nor by differences in beliefs about idea quality. Our results seem most consistent with a preference for promoting and rewarding group members in a way that conforms with gender norms.

Key Words: gender differences, stereotypes, teams, economic experiments

JEL Classifications: C90, J16, J71

[1] We thank Tiffany Chu, Manuela Collis, Kyra Frye, Marema Gaye, and John-Henry Pezzuto for excellent research assistance. We also wish to thank Kristin Butcher, Lucas Coffman, Michael Luca, Corinne Low, Deepak Maholtra, Anya Samek, G. Kartini Shastry, and the participants of the Wellesley College Economics Research Seminar, the Resource Economics Department seminar at the University of Massachusetts Amherst, the Markets, Public Policy, & Law Department Seminar at the Boston University Questrom School of Business, and the session "Interventions to Close Gender Gaps – What Works and What Can Backfire" at the 2020 ASSA meetings and for helpful comments. Coffman gratefully acknowledges financial support from Harvard Business School, and Flikkema and Shurchkov gratefully acknowledge financial support from Wellesley College Faculty Awards. All remaining errors are our own.

[2] Corresponding author. Negotiations, Organizations, & Markets Unit, Harvard Business School, Soldiers Field Boston, MA, USA, kcoffman@hbs.edu

[3] Graduated from Wellesley College in 2017, currently in the private sector; cflikkem@wellesley.edu

[4] Corresponding author. Department of Economics, Wellesley College, 106 Central St., Wellesley, MA, USA, olga.shurchkov@wellesley.edu

I. INTRODUCTION

Across a variety of careers, professional success requires an ability to voice and advocate for ideas in team decision-making. In this paper, we explore gender differences in the ways in which men and women communicate in team decision-making problems, and how groups decide which members to promote, reward, and recognize. We ask whether there are differences in the propensity of men and women to promote themselves and their ideas in these contexts, and whether they are equally likely to be rewarded for their ideas by others.

Although today women make up more than half of the US labor force and earn almost 60% of advanced degrees, they are not represented proportionally at the highest levels of many professions (Catalyst 2020). The gender gap in representation, as well as in earnings, is particularly large in professions dominated by men and perceived to be stereotypically male-oriented, such as finance (Bertrand et al 2010, Goldin et al 2017) and STEM (Michelmore and Sassler 2016). A large body of research has investigated how differences in preferences and beliefs contribute to these gaps (see Niederle 2016 and Shurchkov and Eckel 2018 for surveys).

One strand of work has focused on differences in the willingness to contribute ideas in group settings. Coffman (2014) documents that women are less willing to contribute ideas in stereotypically male-typed domains, and Bordalo et al (2019) and Chen and Houser (2019) find that these effects are stronger in mixed-gender groups where gender is known. Similarly, Born et al (2018) find that women are less willing to be the leader in a group decision-making task, particularly when the team is majority male. There is also evidence that women are less likely to receive credit for their contributions. Sarsons (2017) finds that female economists who co-author with men receive less credit for joint work in terms of tenure probability, and Isaksson (2018) finds that women claim less credit for team's successes in a controlled laboratory experiment.

This literature suggests that gender stereotypes may play an important role in understanding how teams discuss, decide on, and give credit for ideas. We build on this prior work by designing a controlled in-person laboratory experiment that utilizes free form chat among group members. In this way, we take an important step toward studying real world environments of interest, where "speaking up" and advocating for oneself happens in natural language. In our environment, teams brainstorm answers to questions that vary according to the gender stereotype of the topic involved (the perceived "maleness" of the question). Our first contribution is methodological: the novel "*Family Feud*" type task allows for greater subjectivity in the

"correctness" of different ideas. Furthermore, unlike the tasks used in previous laboratory studies where there is only one correct answer, our task admits multiple possible answers, some better than others. This creates a setting where ideas can be contributed, discussed, and debated by teams via free-form chat. Thus, the contribution of ideas, in our setting, more closely mirrors real-life decision-making environments as compared to the more structured experimental paradigms of Coffman (2014), Bordalo et al (2019), and Chen and Houser (2019). Past literature in other decision-making problems suggests that free-form communication can produce meaningfully different results than more structured paradigms (Charness and Dufwenberg 2010).

After groups discuss their ideas via chat, each member provides an incentivized ranking of everyone in the group, indicating who they would most (and least) like to submit an answer on behalf of the group. Individuals who are selected to answer on behalf of the group have the responsibility of aggregating the group discussion into a single group answer that determines each member's pay. These selected leaders are also rewarded with additional compensation. Our focus is on how men and women self-promote (i.e., how they rank themselves), and how they are ranked by others. Together, these choices determine who is selected to represent the group.

We compare conversations and decisions across two between-subjects treatments that vary whether gender is revealed to fellow group members. Differences in how contributions are made, valued, and rewarded across these two chat treatments would suggest an important role for gender-related biases. We also include within-subject control treatments that turn off the free form communication that characterizes our main chat treatments. This allows us to investigate the role of communication in reducing, or exacerbating, gender biases.

We develop a simple theoretical framework that guides our empirical analysis. We analyze an individual's decision about how to rank herself and her fellow group members as a function of her beliefs about the quality of the answer each group member would submit if selected to lead. We hypothesize that three forms of gender bias are plausible in our setting. First, we could observe bias against women, where, conditional on the true quality of the ideas they would submit, women are less likely to be selected to represent the group compared to men, independent of the question domain or group composition. Second, we could observe stereotyping bias, where individuals are less likely to be selected in more gender incongruent areas (for example, more male-typed areas for women). Finally, we could observe minority-status bias, where individuals are less likely to be selected when they are in the gender minority (for example, the only woman in a group). We

discuss how these patterns could be driven by (i) biased beliefs, both about own ability and about others' abilities, and/or (ii) preferences, such as a taste for conforming to perceived gender norms.

When group members know each other's gender, we find evidence of two of these gender-related biases. While, overall, women are no less likely to be selected to represent the group as compared to men, gender stereotypes play an important role in which individuals are recognized and rewarded by the group. Conditional on the quality of their answers, individuals are significantly more likely to be chosen as the group representative in more gender congruent domains. In our setting, this is primarily driven by discriminatory behavior (stereotyping of others), rather than self-stereotyping. We also see some evidence of minority-status bias: individuals are directionally less likely to be chosen to represent the group when they are in the minority. This is driven entirely by differences in self-promotion: individuals rank themselves much more favorably when they are in the gender majority than when they are in the minority. Thus, gender-related biases play a significant role in shaping which ideas and individuals are rewarded when gender is known. In comparison, there are no gaps of these types in the unknown-gender treatment.

We then turn our attention to unpacking the drivers of the gender biases we observe. We use control treatments, that turn off free-form communication, to try to understand the role that communication plays in contributing to these patterns. Past studies have shown in different contexts that free-form communication can be beneficial from the perspective of efficiency and earnings (Brandts, Charness, and Ellman 2015; Dugar and Shahriar 2018). We contribute to this literature by asking whether free-form natural communication has implications for gender biases in this context. In our control treatments, we remove the opportunity for participants to chat. Interestingly, when we restrict communication in these control treatments, gender gaps are eliminated even when gender is known. Thus, an important contribution of our work is to show that natural language communication seems to exacerbate gender gaps and reliance on gender stereotypes, and to begin to unpack why.

Given the centrality of free form communication in generating gender biases, we analyze the conversation data to provide further insights into the team decision-making process. We have third-party external evaluators ("coders") read chat transcripts and provide their assessments of each group member across a variety of dimensions. We vary, across coder, whether or not they do so blinded to the gender of the participants. When blinded to conversationalist gender, coders

perceive female group members as significantly more competent, more assertive, and warmer than the male group members. But, when gender information is available to coders, these patterns are reversed. Simply knowing that a conversation member is female has a significant impact on perceptions, with women's contributions being viewed significantly less favorably across a variety of dimensions when their gender is known. On the other hand, perceptions of men are the same regardless of whether their gender is known.

We explore whether these perceived differences in assertiveness, competence, and warmth help to explain biases in the selection of group representatives. That is, we ask whether the fact that individuals are more likely to be chosen to represent the group in more gender congruent domains can be explained by the fact that individuals are perceived (at least by our coders) as being more assertive or more competent in those domains. While we do find that groups are more likely to select members who are perceived as more competent, differences in these characteristics do not explain the gender-related biases in who is selected as group representative. That is, even conditional on having the same answer quality, making similar contributions to the group, and being perceived as behaving similarly assertively, competently, and warmly, individuals are still less likely to be chosen to represent the group in gender incongruent domains.

In our final step, we try to disentangle the roles for beliefs and preferences in contributing to the gender biases we observe. The most obvious explanation for the patterns we observe are biases in beliefs about the abilities of different group members to submit a high quality answer for the group. We use belief data elicited from our participants after the conversations, at the time of their ranking decisions, to explore this explanation. We show that while beliefs (both about self and others) are predictive of the likelihood of being chosen as the group representative, beliefs do not explain a significant portion of the gender-related biases we observe. Even controlling for beliefs about the quality of the answer each member would submit, individuals are significantly more likely to be chosen as group representative in more gender congruent domains. And, again conditional on beliefs they hold about themselves and others, individuals are more likely to self-promote when they are in the majority rather than the minority in terms of the gender composition of the group. While we can only speculate, this could reflect a perhaps implicit preference for conforming with perceived gender norms.

Our results are consistent with a growing literature showing the importance of stereotypes for economic outcomes. Beliefs and behavior seem to depend upon the gender-type of the domain

at hand. For example, Shurchkov (2012), Dreber et al. (2014), and Grosse et al. (2014) show that gender gaps in the willingness to compete become substantially smaller and insignificant in the context of a more female-typed task as compared to the more stereotypically male-typed task used by Niederle and Vesterlund (2007). Similarly, Hernandez-Arenaz (2018) finds that men who perceive a task as more male-oriented have more optimistic self-assessments of ability and are more likely to enter a high-paying tournament. Previous studies have also shown that female decision-makers are more likely to act in a gender-congruent way when their gender would be observable (see for example, Charness and Rustichini 2011 for the decision to cooperate and Shurchkov and van Geen 2019 for the decision to assign competitive incentives to workers). Public observability in the presence of gender stereotypes has also been shown to significantly decrease women's willingness to lead (Alan et al 2017, Born et al 2020), willingness to compete (Buser et al 2017), and willingness to express ambition (Bursztyn et al. 2017).

The complexity of our experimental environment, which includes a subjective task, group interaction, natural language conversation, and several treatment variations, generates a wealth of data, well-suited to answering a variety of interesting, new questions about these types of gender biases. We use a simple experimental model to discipline our empirical analysis. In line with past work, we find important roles for the observability of gender, the gender-type of the domain, and the gender composition of the group in shaping decisions. In a single, unified setting, we document the different influences of these factors on both decisions related to self-promotion and evaluations of others. Using a variety of sources and types of data, we show that these gender biases are not well-explained by differences in conversational behavior or beliefs about answer quality, suggesting a need to further understand the preferences that may be at work in these types of group decision-making problems.

II. THE EXPERIMENT

II.A. THE TASK

Participants in our experiment play multiple rounds of a *Family Feud* style task.[5] *Family Feud* is a popular gameshow in which teams attempt to guess how respondents in a survey answered different questions. To our knowledge our study is the first to use a modified *Family Feud* game

[5] Questions were selected from the database at http://familyfeudfriends.arjdesigns.com/ For more information about the game show *Family Feud* see, for example, https://www.thoughtco.com/family-feud-brief-overview-1396911

in an economic experiment. The task was chosen to mirror many of the real-world properties of group decision problems. In this task, there are many good answers to most questions. Some answers are better than others, but there is room for disagreement. This feature mimics the real-world properties of brainstorming that is under-studied in real-effort task experiments. The points that are ultimately earned by the group depend upon the answers given by the previous survey respondents. Individuals could play this game independently, but there is room to learn from and debate with others. Thus, the task combines the two desirable features for our purposes: a high degree of subjectivity, and still a clear scoring system which admits no ambiguity as to how answers are ranked.

Our version of *Family Feud* works as follows. Individuals are shown a question, and the goal is to guess an answer to the question that would be frequently given by the previous survey respondents. Specifically, the *Family Feud* questions we source have been previously shown to a 100-person survey panel, who each gave an answer to the question. These survey panel answers generate the scoring system for the game. The number of points a given answer is worth is equal to the number of survey respondents who gave that particular answer. Thus, players in our experiment should aim to provide answers that were popular among the survey respondents, and hence are worth more points.[6] Consider the example below which we presented to subjects in the instructions for practice.

Example: "Name a word a judge might yell out during a tennis match"

Answers	***Points***
Fault	*25*
Foul	*17*
Love	*14*
Out	*10*
Order	*6*
Net	*4*
Point	*3*

Here, "fault" receives the most points because 25 out of 100 surveyed individuals stated this as their answer to the given question. However, "foul" or "love" are still valuable answers, as they yield some points, albeit less than the top answer. Only answers that receive two or more survey responses count for points. Note that, for scoring purposes, it does *not* matter how many participants *in our experiment* gave a particular answer; the points were simply based upon these

[6] Each point is worth $1 in compensation. We provide full details on the incentives in Section II.C.

100-person survey panels constructed by *Family Feud*. Our participants were informed of this scoring system, so that they understood that "best" answers were those most popular on the survey and not necessarily those which they felt were the most correct or the most inventive.[7]

In summer of 2017, we conducted a pilot on Amazon Mechanical Turk (AMT) to determine the most appropriate *Family Feud* questions for the purposes of our study (see details and a complete set of questions in Appendix G). The goal was to determine common answers to each question (so that we could program our experiment to accept common variants of each answer), and to understand the gender stereotype associated with each question. Within the pilot, each AMT participant provided several answers to a subset of questions drawn from 20 different *Family Feud* questions. And, they provided their perception of the gender stereotype for each question, indicating for each question on a -1 (strongly favors women) to 1 (strongly favors men) scale whether they believed men or women would be better at answering that particular question. Using this data, we selected 8 questions: four perceived as female-typed and four perceived as male-typed. We use these 8 questions in our main experiment, randomly assigning one to each round of the experiment at the session level.

The extent to which a question is perceived to carry a male-typed stereotype, as perceived by these AMT pilot participants, informs one of our main variables in the subsequent analysis. We refer to this as the "maleness" index, which ranges from -0.57 (the average slider scale rating of the most female-typed question) to 0.51 (the average slider scale rating of the most male-typed question). We are interested in how behavior responds to the extent to which a given question is gender congruent: more male-typed for men, or more female-typed for women. Thus, we will predict outcomes from the "gender congruence" of a question: for men, this is exactly the maleness index of the question, and for women, it is the maleness index re-signed (-1*maleness). This allows us to ask, for any individual, how behavior changes as the question becomes more or less gender congruent in terms of its associated stereotype. The set of questions we use in our main experiment, along with associated maleness score, is below.

[7] Subjects were also cautioned to check the spelling of their submissions, since misspelled answers could result in a score of zero points. In practice, we coded the experimental program to accept common misspellings and common variants of each possible answer. These were sourced through an online pilot. But, we still wanted to caution participants that we could not guarantee that misspellings would be recognized.

Question	Avg. Maleness
Name an item of clothing that you should not wash in the washing machine.	-0.57
Name a city or state that has a name a woman might have.	-0.40
Name a reason your eyes might water/have tears.	-0.24
Name a drink or good that can be consumed either hot or cold.	-0.13
Give me a word or phrase that contains the word "bar."	0.11
Name a sport in which the competitors wear funny-looking shoes.	0.31
Name something a fire-fighter doesn't want to be without.	0.32
Name something men think is manly.	0.51

II.B. EXPERIMENTAL DESIGN

In our experiment, participants play repeated rounds of the *Family Feud* game, each time in a new group. Each session of the experiment consisted of two parts, each containing four rounds of interactions. Each round uses a unique one of the eight *Family Feud* questions. Within a given part, participants were randomly re-matched in groups of three for each round, using stranger matching. All interaction took place via private computer terminals.

Figure 1 summarizes the stages and the flow of the experiment within each round. Each round began with a "*pre-group*" stage where participants had 15 seconds to view the question and 30 seconds to submit an individual answer. After submitting the answer, subjects were asked: "*On a scale of 1-10, please indicate how confident you feel about your ability to submit a high-scoring answer to this specific question*." This gives us a pre-group measure of individual ability and individual confidence.

Next, subjects entered the "*group*" stage where they could chat over the computer interface for 60 seconds with each other. This gave groups a chance to volunteer, debate, and discuss different answers. At the end of the chat, participants view a chat transcript. Participants then ranked each member of their group, including themselves, from 1 – 3, where 1 indicated the person they would most want to answer on behalf of the group, i.e. be "the group representative." Within each group, we randomly chose one participant whose ranking would then determine the actual group representative (random dictatorship). We used that randomly-selected participant's ranking to probabilistically select a group representative: the person they ranked first had a 60% chance of being the group representative; the person they ranked second had a 30% chance; the person they

ranked third had a 10% chance. In this way, we incentivize each group member to provide a complete ranking of the entire group, as any participant could be chosen to determine the group representative, and their full ranking is relevant for this determination. Alongside this ranking, each group member also provided a subjective "confidence" of each group member's ability to provide a high scoring answer to that question (again on a 1 – 10 scale). This is our measure of the beliefs each respondent holds about the likely quality of each group member's answer, should that group member be selected as the group representative

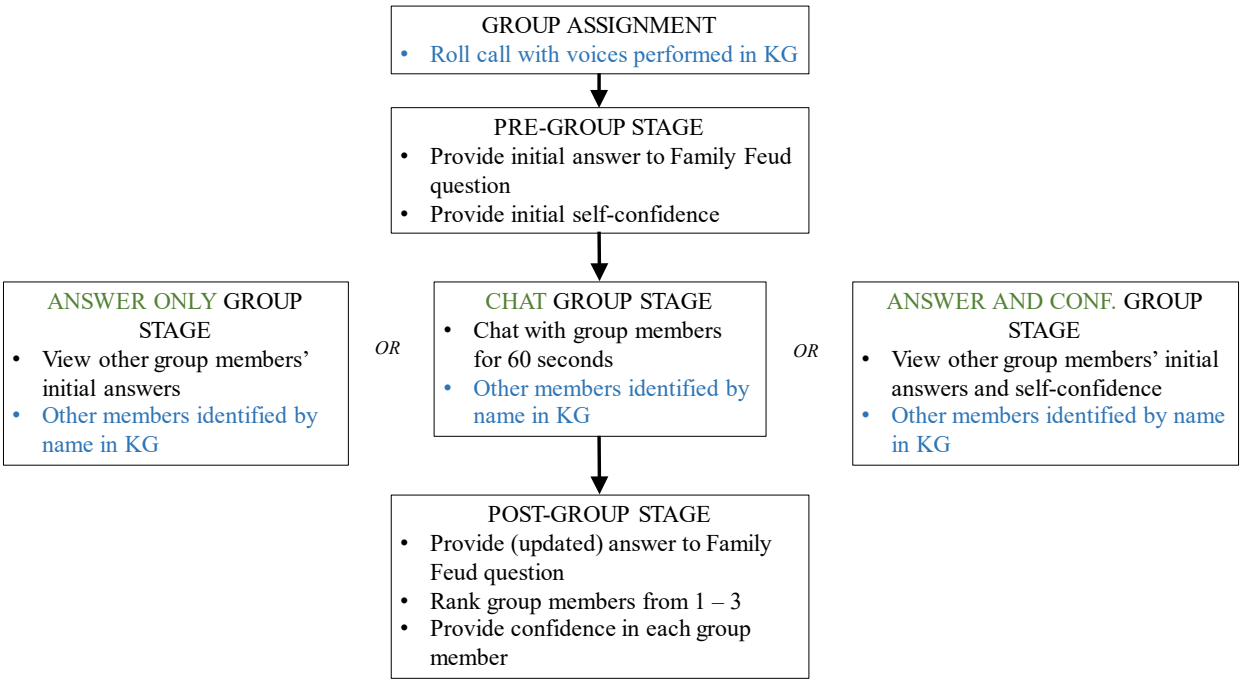

Figure 1: Stages of Each Round

The "group representative" is important, both because he or she determines which answer will be submitted on behalf of the group, aggregating the group's discussion into a single, collective outcome, and also because he or she receives a material incentive for being selected to serve in this capacity – a bonus of $2. In this way, being chosen as the group representative carries responsibility and increased compensation, reflecting the features of being recognized for one's contributions and being promoted to positions of leadership outside of the laboratory.

Finally, there was a "post-group" stage where subjects again submitted individual answers to the same *Family Feud* question. Subjects knew that, if they were selected as the "group representative," this would be the answer submitted on their behalf. This also allows us to document how individual answers were influenced by the group discussion, and to provide the counterfactual of how each individual would perform if chosen as the group representative.

Our primary treatment variation is whether or not gender information is made available to participants. In the unknown-gender treatment (UG), participants were identified in each round by a randomly-generated ID number. In the known-gender treatment (KG), we revealed gender to participants. We did this in two ways. First, we had group members provide their first name at the beginning of the first round of the treatment. They were encouraged to use their real name, but participants were able to enter any name they wished.[8] This name was then used throughout the part to identify them to their fellow group members during their computer interactions. Second, we did a verbal roll call, in which groups were announced out loud in each round, and each member of the group was called by their provided name and asked to respond "here". In this way, the rest of their group members were likely to identify their gender, even if their name was ambiguous (as in Bordalo et al 2019). Note that because the laboratory was equipped with partitions and participants remained seated at their private terminals, participants were unlikely to view their fellow group members during this process. Thus, while they learn their group member's name and hear their voice, they do not see what they look like.[9]

In addition to the varying whether gender was known, we varied the extent to which group members could communicate with one another. Beyond our main chat treatment, where subjects freely communicated via computerized chat during the group stage, we have two control treatments aimed at understanding the mechanisms at work. The control treatments parallel the chat treatments, but eliminate the opportunity to chat. In the answer only control treatment, we simply display the answers submitted in the pre-group stage during the group stage. In the answer plus confidence treatment, we display both the answers submitted and the self-confidence rating from the pre-group stage during the group stage (more detail on these treatments is provided after

[8] 91% of participants report in the post-experiment questionnaire that they used their real name. We use this indicator as a control variable in our specifications.

[9] One might ask whether participants were likely to know other individuals in their session. We ask participants in the post-survey questionnaire whether they recognized anyone in their sessions; 85% of participants report they did not recognize anyone in their session. We then asked, if they did recognize someone, whether that knowledge changed any of their decisions: 92% of participants report that it did not.

our main analysis, in Section IV.B). In each session, subjects participated in exactly two treatments, one in each part. In every case, one of these treatments was a Known Gender (KG) treatment and the other was an Unknown Gender (UG) treatment, and at most one was a chat treatment. Figure 2 summarizes the way in which subjects were randomized into treatments.

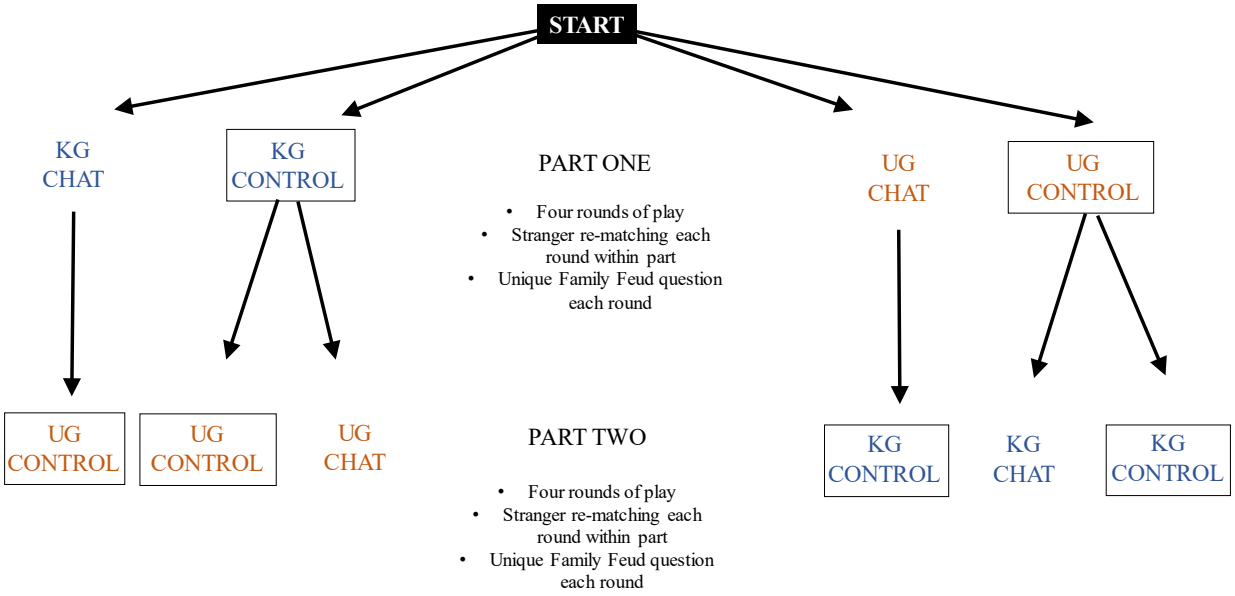

Figure 2: Randomization into Experimental Treatments (KG = Known Gender; UG = Unknown Gender)

II.C. Incentives and Logistics

One round was randomly selected for payment at the end of the experiment. Participants were paid based upon one of three submissions in that round: there was a 10% chance they were paid for their individual answer in pre-group stage, an 80% chance they were paid for the group answer given by the selected group representative, and a 10% chance they were paid for their individual answer in the post-group stage. Participants were paid $1 for every point earned by the randomly-selected answer. In addition, the person selected as the "group representative" in the randomly-selected round received a bonus payment of $2.

After signing the informed consent form, participants were seated at individual computer terminals. Subjects received written, oral, and on-screen instructions programmed using the standard zTree software package (Fischbacher 2007). Participants were encouraged to ask questions in private if they did not understand these instructions, but communication between subjects was disallowed other than when instructed. Subjects only received the instructions

relevant to the immediate part of the experiment (Part 1 or 2). At the end of the experiment, subjects were informed about their performance and payment and filled out a post-experiment questionnaire with demographic questions (instructions and questionnaire are available in the online Appendices H1 and H2). Each session of the experiment lasted approximately one hour. Subjects were paid in cash and in private by the experimenters. Mean payment across all sessions, including the show-up fee, was equal to $26.48.

One pilot session (data excluded from analysis) and 19 sessions of the experiment were conducted at the Computer Lab for Experimental Research (CLER) at Harvard Business School (HBS) between September 2017—May 2018. Our ex ante plan was to obtain approximately 100 observations in each treatment cell, and we stopped running sessions when we hit that target. In total, we have 297 participants, each of whom participated in two treatments. In our primary analysis, we focus on our main chat treatments: 207 subjects participated in our main chat treatments, 105 in the Known Gender version and 102 in the Unknown Gender version (Table 1). Ex post, it is clear that we are under-powered to explore some questions that would potentially be of interest, particularly those that consider interactions of variables of interest. We are mindful of statistical power considerations in our analysis, expressing caution when reporting specifications that rely on small numbers of observations per cell. We hope that our results spark future work that is more well-powered to consider interesting research questions that we omit.

Table 1. Experimental Treatments

	Known Gender	**Unknown Gender**
Main condition: chat via computer	**105**	**102**
Control: answer + confidence observable	87	87
Control: only answer observable	105	108
Total subjects	297	297

Notes: Each subject was assigned to a known gender treatment in one part of the experiment, and an unknown gender treatment in the other part of the experiment. The total number of unique subjects is 297.

III. EXPERIMENTAL MODEL AND HYPOTHESES

In this section, we present a framework for understanding the experimental environment our participants face. Our key outcome variable is the likelihood of being chosen as the group representative, which is determined by the rankings provided by each group member. Our model analyzes an individual's decision of how to rank her group members. In our simple model, the participant's beliefs about the quality of answer each group member would submit as group representative determine their ranking decisions. Preferences can also play a role. We outline this formally below and explain how the model informs our empirical approach.

We assume that the goal for each participant in determining her rankings is to maximize her compensation for the group stage. She can earn compensation through a higher quality answer being submitted for the group (where each point earns \$1) and through being chosen as the group representative (where the group representative earns an additional \$2 bonus). Let $p_i(x_i)$ be the number of points that individual *i* believes her own answer that she would submit as the group representative, x_i, is worth. Let $p_i(x_j)$ be the number of points that individual *i* believes participant *j*'s group answer would be worth, and $p_i(x_k)$ be the number of points that individual *i* believes participant *k*'s group answer would be worth.

Without loss of generality, assume that $p_i(x_j) \geq p_i(x_k)$; that is, participant *i* believes that player *j's* submission would generate weakly more points than that of *k*. So, participant *i* has three plausible rankings to choose from: ranking herself first (expression 1), ranking herself second behind player *j* (expression 2), or ranking herself third behind *j* and *k* (expression 3).

$$0.6\,(p_i(x_i) + 2) + 0.3p_i(x_j) + 0.1p_i(x_k) \quad (1)$$

$$0.6\,p_i(x_j) + 0.3(p_i(x_i) + 2) + 0.1p_i(x_k) \quad (2)$$

$$0.6\,p_i(x_j) + 0.3p_i(x_k) + 0.1(p_i(x_i) + 2) \quad (3)$$

Here, 0.6/0.3/ 0.1 are the probabilities with which a group member ranked first/second/third becomes the group representative. Taken together, these equations imply that participant *i* will rank herself ahead of another participant, $n \in \{j, k\}$, if $p_i(x_i) > p_i(x_n) - 2$.

That is, a participant will rank herself first if she believes that the answer she will submit as group representative would score no less than 2 points worse than the answer that would be

submitted by the next best group member. If, all else equal, a participant believes that every group member will submit the same answer, or if she is confident that her answer is the best among those likely to be submitted, she will rank herself first. However, if a participant is not confident in her answer and believes that participant *j* will submit a sufficiently better answer, she will instead rank participant *j* first. In fact, if she feels both *j* and *k* will submit answers sufficiently better than her own (at least 2 points better), she will rank herself last.

Notice that these ranking decisions embed two distinct components. The first component is how an individual chooses to rank herself: first, second, or third. This ranking of self, which we will call self-promotion, depends upon the relative value of $p_i(x_i)$ compared to $p_i(x_j)$ and $p_i(x_k)$. Self-promotion decisions thus depend both on beliefs about self, and on beliefs about others. The second component is how an individual chooses to rank the other two members of her group relative to each other: does she rank *j* above *k*. Here, beliefs about self are irrelevant; it is simply *i*'s belief about the other two group members, $p_i(x_j)$ and $p_i(x_k)$, that guides this decision.

Our key outcome variable is the total likelihood of any individual being chosen as the group representative: the natural aggregation of (1) self-promotion decisions (how each group member chooses to rank themselves) and (2) decisions about the evaluations of others (how each group member ranks the remaining two group members relative to each other). In our empirical analysis, we will also separately consider both self-promotion decisions and the rankings of others, exploring how each contributes to the patterns observed in the total likelihood of being selected.

Our first key assertion is that beliefs, $p_i(x_i)$, $p_i(x_j)$ and $p_i(x_k)$, may be biased. In particular, when gender is known, there are three plausible biases that could distort beliefs: gender bias, gender stereotyping, and minority status bias. We work through each of these below.

We first consider gender bias: conditional on true idea quality, women's ideas may be perceived as lower quality than men's ideas. Note that gender bias could show up both in assessments of self, $p_i(x_i)$, and in assessments of others, $p_i(x_j)$ relative to $p_i(x_k)$. Gender bias in beliefs about self, i.e. gender differences in self-confidence, would lead men and women with similar ideas to have different perceptions of those ideas. Holding all else equal, this could lead to gender gaps in how favorably men and women rank themselves.[10] Gender bias could also impact beliefs about others. That is, consider two group members, a male player *j* and a female player *k*,

[10] This gender bias in self-confidence could show up not only when gender is known, but also in the unknown gender treatment.

whose ideas are worth the same number of points. Gender bias would imply $p_i(x_j) > p_i(x_k)$, leading to a greater likelihood of men being ranked ahead of women by others. Together, gender bias, through its impact on beliefs about self and others, would lower the total likelihood of women being selected as the group representative.

Hypothesis 1: When gender is known, we will observe **gender bias:** on average, women will be less likely than men to be selected as the group representative conditional on true answer quality.

The second type of bias we may observe is gender stereotyping. Again, gender stereotyping could impact both beliefs about self and beliefs about others, when gender is known. Under gender self-stereotyping, we expect that $p_i(x_i)$ would increase with the extent to which the question is drawn from a domain that is gender congruent (more male-typed for men, more female-typed for women), even holding true answer quality, x_i, fixed. In our known gender treatments, stereotyping can also impact $p_i(x_n)$, with beliefs about another player's abilities again increasing with the extent to which the question is drawn from a domain congruent with the other person's gender. Together, these can impact both self-promotion decisions (whether $p_i(x_i) > p_i(x_n) - 2$), as well as the evaluations of others (whether $p_i(x_j) > p_i(x_k)$). Through both of these channels, we expect that men (women) will be more likely to be selected as the group representative for more male-typed (female-typed) questions, holding idea quality fixed. This effect should be stronger when gender is known, because in this case stereotyping of others, not just of self, is possible.

Given these dynamics, it is clear that the gender composition of the group could also be an important factor in the known gender treatment. If players *i* and *j* are the same gender, stereotyping moves both $p_i(x_i)$ and $p_i(x_j)$ in the same direction. In this case, the relative ranking should be less dependent on the gender stereotype of the question. However, when players *i* and *j* are of different genders, stereotypes drive a wedge between $p_i(x_i)$ and $p_i(x_j)$. Thus, we expect stronger effects of stereotyping in mixed gender groups, compared to single gender groups.

Hypothesis 2: When gender is known, we will observe **gender stereotyping**: participants will be more likely to be selected as group representative in gender-congruent domains (questions drawn

from male-typed domains for men, female-typed domains for women). This effect will be stronger in mixed-gender groups.

Past literature suggests a final way in which beliefs may be biased: minority status bias. Previous evidence on team decision-making suggests that the gender composition of the group may be relevant for beliefs about self. Holding all else fixed, individuals may feel more confident when they are in the majority rather than the minority (see, for instance, Born et al (2020) and Shan (2020)). Stated in our framework, we predict that $p_i(x_i)$ increases in the share of group members that have the same gender as participant i. As $p_i(x_i)$ increases, rankings of oneself improve because it becomes more likely that $p_i(x_i) > p_i(x_n) - 2$. This should happen only when information on group composition is available – i.e. when gender is known.

Hypothesis 3: When gender is known, we will observe **minority status bias**: participants will be more likely to be selected as the group representative as the share of same gender group members increases. This will operate through self-promotion decisions.

This simple framework guides our analysis below. In particular, we will test these three hypotheses, predicting the probability of being chosen as the group representative from three key features. First, in line with Hypothesis 1, we ask whether women are less likely to be selected as the group representative, independent of the gender stereotype of the question or group composition. Second, in line with Hypothesis 2, we ask whether the gender congruence of the question has predictive power for the likelihood of a given member being selected as a group representative, and whether this effect is stronger when the group is mixed gender. Third, in line with Hypothesis 3, we ask whether the share of same gender group members positively predicts the likelihood of being selected as the group representative. If these patterns are indeed driven by gender biases, they should be more prominent in the known gender treatment than in the unknown gender treatment.

Up to this point, we have formulated each of these gender-related biases as a product of belief-based, or statistical, discrimination. But, it may be the case that preferences, or taste-based discrimination, also play a role in these decisions. In fact, for each of the three hypotheses laid out above, similar patterns could be driven by tastes, in addition to or instead of, biased beliefs.

For instance, gender bias could be driven by taste-based discrimination: participant *i* could incur a cost, $c>0$, for having a woman chosen as the group representative. Such a cost would decrease the chances of women being selected as the group representative, regardless of beliefs. While there is not clear past evidence suggestive of it, it is also plausible that the patterns of gender stereotyping or minority status bias could also be driven by tastes. That is, stereotyping may impact rankings because of preferences for choosing group representatives in a way that conforms to gender norms, or preferences for asserting oneself more in more gender congruent domains, independent of beliefs about the quality of an answer. Similarly, group members may be more willing to assert themselves when in the majority because of a distaste for being selected as a minority group member (for example, due to aversion to "tokenism"), or a fear of being judged more critically for a negative outcome. In our analysis, we will attempt to disentangle the role of beliefs from preferences. We will examine the extent to which gender bias, gender stereotyping, and minority status bias, if observed, are well-explained by participants' self-reported beliefs about the quality of each group member's answer.

IV. RESULTS

Appendix Tables A1 and A2 provide summary statistics of our participants. On average, we find no statistically significant differences in any of the demographic characteristics by gender, other than that our male subjects are significantly more likely to identify as Hispanic (12%) than our female subjects (4.7%). Men and women do not significantly differ in average performance: in the pre-group stage, men earn 14.1 points and women earn 13.1 points on average for their answers (t-test p-value of 0.37). In Appendix Table A3, we confirm balance on demographics across the chat and the control treatments (recall that all subjects participate in both our known gender and unknown gender treatments). In order to improve the precision of our estimates, we control for individual characteristics in our main analysis. (Appendix B confirms that our results are robust to omitting these controls.)

IV. A. BIASES IN THE LIKELIHOOD OF BEING CHOSEN AS GROUP REPRESENTATIVE

Our key outcome variable is the total likelihood with which an individual is chosen as the group representative for a given question. This is a function of the rankings of each of the three group members. In the tables below, this variable is presented on a 0 – 100 scale.

Following our model, we ask whether there are gender-related biases in the total likelihood of being selected to represent the group. We test for gender bias (Hypothesis 1) by including an indicator for the individual being female, asking whether, overall, women are less likely to be selected than men. We test for gender stereotyping (Hypothesis 2), asking whether individuals are more likely to be selected in more gender congruent questions. To do so, we include a variable that proxies for how gender congruent the question is for the individual (for men, this is simply the maleness index for the question; for women, we reverse the sign on the maleness index). Third, we test for minority status bias (Hypothesis 3), asking whether individuals are more likely to be chosen when they are in the majority, rather than in the minority. To do so, we construct an own gender share variable, which represents the share of the other two group members who match the individual's gender. For a man in a group with two women, this variable takes a value of 0/2; for a man in a group with a woman and a man, it takes ½; and for a man in a group with two men, it takes 2/2.

We use a linear probability model to predict the total likelihood that a given member is chosen as group representative, splitting the analysis by treatment (KG and UG).[11] We are interested in the likelihood of individual i being chosen conditional on the quality of the answer she would submit (x_i in our experimental framework). So, in our analysis, we control for the individual's post-group stage answer; that is, the answer she would submit for the group. We include a number of other controls, to account for differences in ability entering the group stage and demographics that may vary across individuals.[12] We cluster standard errors at both the individual and the group level.

---

[11] Ordered probit specifications deliver similar results (see online Appendix B).

[12] We proxy for baseline ability by controlling for the quality of individual pre-group answer (points her pre-group answer would earn), the quality of her individual pre-group answer relative to the mean quality of individual pre-group answers in her group (i.e. the difference between points that would be earned by her given answer less average points for individual pre-group answers in her group), and the quality of the individual answer relative to the best pre-group answer in her group (i.e. the difference between points that would be earned by her given answer less points for highest-scoring individual pre-group answer in her group). We also control for part and round fixed effects, as well as demographic characteristics.

Table 2 presents our main results. We find no evidence in support of Hypothesis 1: conditional on the quality of their ideas, women are no less likely to be selected as the group representative in both the KG and UG treatments. We do find evidence of gender stereotyping, consistent with Hypothesis 2. In particular, in the known gender treatments, individuals are significantly more likely to be chosen as the group representative when the question is more gender congruent (Column 1). We estimate that moving from the least congruent question to the most congruent question increases the chances that an individual is chosen by approximately 4 percentage points.[13,14] There is no evidence for a similar pattern in the unknown gender treatment (Column 2). When we use an interacted model, we see that the effect of stereotyping is larger in the KG treatment than the UG treatment (Column 3, $p<0.10$).

We find directional evidence in favor of Hypothesis 3. Overall, we estimate that, as the same gender share of group members increases, individuals are directionally more likely to be selected as the group representative. While this effect is larger in the KG treatment than in the UG treatment, it is not significant in either specification, and we identify no significant difference in this effect between the treatments (Column 3).

As we outlined in our experimental model, we expect that gender stereotyping, if present, may be particularly pronounced in mixed-gender groups. In Columns 4 – 6 of Table 2, we restrict our attention to mixed-gender groups and see evidence consistent with this prediction. In mixed-gender groups, we see a significant impact of stereotyping ($p<0.05$)in the KG treatment, but not in the UG treatment. Comparing Columns 1 and 4, we estimate that the effect of stereotyping is approximately 30 percent larger in mixed gender groups as compared to the full sample.

[13] We estimate this by differencing the maleness value for the most and least male-stereotyped question and multiplying by the coefficient on gender stereotype: (0.51 + 0.57)*3.7.

[14] In a model that interacts the female dummy with gender congruence, we find no evidence that there is a gender difference in the extent to which gender congruence predicts the likelihood of being chosen.

Table 2. The Determinants of the Probability of Being Chosen as the Group Representative Post-Group Interaction in the Chat Treatment

Sample	All Groups			Mixed-Gender Groups		
	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.021	1.708	1.650	0.0507	2.557*	2.455*
	(1.240)	(1.205)	(1.196)	(1.459)	(1.414)	(1.399)
Gender Congruence of Question	3.715**	-0.716	-0.975	4.796**	-0.615	-0.811
	(1.820)	(1.875)	(1.830)	(2.180)	(2.211)	(2.195)
Own Gender Share in Group	2.748	1.038	1.066	5.829*	3.319	3.733
	(1.689)	(1.616)	(1.594)	(3.048)	(2.771)	(2.816)
KG Treatment			0.501			0.755
			(1.770)			(2.102)
Female x KG			-1.516			-2.115
			(1.691)			(1.992)
Gender Congruence x KG			4.562*			5.409*
			(2.551)			(3.060)
Own Gender Share x KG			1.028			1.566
			(2.301)			(4.105)
Performance Controls	YES	YES	YES	YES	YES	YES
R-squared	0.117	0.139	0.0949	0.143	0.159	0.115
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. Dependent variable mean is 33.33. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Performance controls include points in pre- and post-group stage and the pre-chat distribution that includes difference from maximum group score and difference from average group score. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

In our experiment, two distinct factors influence the total likelihood of being selected as the group representative. First, group representative selection is a function of how each individual ranks themselves, relative to the other two members of the group. One could think of this as the extent to which each individual self-promotes. Second, these likelihoods depend on how each individual evaluates others. That is, for individual *i,* her probability of being chosen also depends on how participant *j* evaluates her relative to participant *k.* As our theoretical framework lays out, gender-related bias has the potential to operate either through differential self-promotion (i.e., differences in how self is ranked across gender, across gender congruence, or across group composition), and/or through discrimination in the relative treatment of the two other group

members. With this in mind, we now break down our analysis, examining our three main hypotheses separately for self-rankings and for rankings of others.

We begin with self-rankings. Note that, perhaps unsurprisingly, the modal action for participants is to rank themselves first – with participants giving themselves the top ranking in 68% of interactions. They rank themselves second in another 19% of interactions, and last in only 12% of interactions. If participants believe all group members are likely to submit similar answers, and they care about the financial reward for being chosen, this finding is in line with the model we outlined in Section III. But, even given this high degree of overall self-promotion, we see interesting patterns within the self-rankings.

Table 3 parallels Table 2, but this time predicting individual *i*'s self-ranking: first (60 percent probability of being selected as group representative), second (30 percent), or third (10 percent), again on a 0 – 100 scale. We include all the same controls, and our three key variables: individual *i*'s gender, the gender congruence of the question for her, and the share of other group members of the same gender group members as her.

The results suggest that, overall, men and women self-promote to a similar extent, both in the KG and UG treatments. That is, we see no significant differences in self-rankings by gender. Similarly, gender stereotypes do not seem to be a central factor in either treatment: individuals are not more likely to rank themselves more favorably in more gender congruent questions. These null results are interesting findings, and at odds with some related work in this area (for instance, Coffman (2014), Chen and Houser (2020), or Bordalo et al (2019) on self-stereotyping). One important explanation could be the role for incentives. In each of these previous papers, group member incentives were entirely aligned: there were no bonus payments associated with being the person who submitted the answer for the group. In our framework, there is a clear and non-trivial financial reward for being the person chosen to submit. It could be that this financial incentive is enough to overcome the impact of self-stereotyping. A second important factor could be the role for deliberation. In our framework, again contrary to those past studies, groups have a chance to hear from each other before selecting who answers for the group. It might be the case that after having the chance to discuss, and to take into account other's answers, there is less of a role for self-stereotyping, in part because individuals can incorporate information and ideas from others in forming their final answer.

Table 3. The Determinants of Self-Rankings Post-Group Interaction in the Chat Treatment

Sample	All Groups			Mixed-Gender Groups		
	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-1.450	1.045	1.038	-0.637	3.105	2.715
	(1.856)	(1.835)	(1.810)	(2.167)	(2.254)	(2.194)
Gender Congruence of Question	2.010	-2.250	-1.455	3.378	-0.930	-0.589
	(2.721)	(2.755)	(2.745)	(3.229)	(3.393)	(3.393)
Own Gender Share in Group	6.330**	1.972	1.988	9.410**	4.699	5.189
	(2.736)	(2.444)	(2.434)	(4.768)	(4.489)	(4.540)
KG Treatment			-2.371			-1.792
			(2.723)			(3.176)
Female x KG			-2.252			-2.830
			(2.519)			(3.031)
Gender Congruence x KG			3.239			3.773
			(3.858)			(4.660)
Own Gender Share x KG			3.443			3.459
			(3.589)			(6.478)
Performance Controls	YES	YES	YES	YES	YES	YES
R-squared	0.119	0.102	0.0835	0.114	0.147	0.0967
Dep. Var. Mean	46.67	49.26	47.95	46.04	49.12	47.50
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Performance controls include points in pre- and post-group stage and the pre-chat distribution that includes difference from maximum group score and difference from average group score. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

We find that self-promotion decisions are sensitive to group composition. In the KG treatment, individuals are more likely to self-promote as the share of same gender group members increases, in line with Hypothesis 3 (Columns 1 and 4). We do not see a significant effect in the UG treatment (Columns 2 and 5); however, we cannot reject that the effects of group composition are of the same magnitude across treatments in an interacted model (Columns 3 and 6).

We now turn our attention to rankings of others: when we take self out of the equation, are there gender-related biases in how others are ranked? We shift our empirical approach to allow for a careful examination of this question. In Table 4, we predict the likelihood with which individual *i* ranks individual *j* above individual *k.* Our focus is on how gender-related biases impact the relative ranking of *j* and *k.* We again use a linear probability model, where the dependent variable

is simply a dummy if *j* is ranked above *k* by individual *i*. We scale this to a 0 – 100 scale, to parallel our previous analysis. Note that for each individual-round observation, we have one independent realization of this variable: the relative ranking of individual *i*'s two remaining group members for that round.

In testing our hypotheses in this setting, we need to adapt our empirical approach. In particular, we need to measure the extent to which each of our key variables (gender, gender congruence of the question, and share of same gender group members) *differs* across *j* and *k.* For Hypothesis 1, we want to ask how the genders of *j* and *k* impact their relative ranking. To get at this, we construct a variable that simply differences the female dummies for *j* and *k*. In a sense, we are capturing how female *j* is relative to *k* (recognizing that these are simply differencing binaries). The coefficient on the difference between female dummies for *j* and *k* allows us to test Hypothesis 1: a negative coefficient suggests discrimination against women.

We take a similar approach to identifying the role for gender stereotypes. We construct the difference in the extent to which a question is gender congruent for *j* and *k.* Suppose *j* and *k* are of the same gender. Then, for any particular question, the gender congruence of that question will be the same. Thus, for same gender pairs, this variable -- the relative gender congruence -- takes 0. When *j* and *k* are of opposite genders, this relative gender congruence variable captures how much more (or less) the stereotype benefits *j* relative to *k*. The coefficient on this relative gender congruence term allows us to test for gender stereotyping (Hypothesis 2) in the ranking of others.

Finally, we use this same differencing construction to capture the extent to which the group composition matches *j*'s gender, relative to *k*'s. We simply subtract the same gender share for *k* from the same gender share for *j*. This allows us to test for Hypothesis 3 in rankings of others, gauging whether minority group members are ranked less favorably (by others) than majority group members.

Again, we are interested in the extent to which these gender biases shape decisions, conditional on the true quality of answers. For that reason, we also control for the relative quality of *j*'s ideas relative to *k*'s. That is, we control for the difference in the number of points associated with the pre- and post-group answers submitted by *j* and *k.* In addition to these key variables, we control for our standard set of controls and the demographic characteristics of the individual providing the ranking (individual *i* in the terminology above).

The results are reported in Table 4. We again find no evidence of overall gender discrimination: if anything, women are more likely than men to be ranked above a fellow group member in both the KG and UG treatments. But, we see a sizable and significant impact of stereotyping in the KG treatment. Consider two group members, one male (*j*) and one female (*k*). We estimate that moving from the most male-typed question (where the relative gender congruence is 1.02) to the least male-typed question (where the relative gender congruence is 1.14), would decrease the likelihood of the man being ranked above the woman by his fellow group member by approximately 26 percentage points (2.16 x 12). This effect is statistically significant with $p<0.05$ in the KG treatment, while the estimated coefficient is close to 0 in the UG treatment. In an interacted model, the coefficient on relative gender congruence interacted with the KG treatment is positive and large, but noisily estimated.

While we find directional evidence that group members in the majority are more likely to be ranked ahead of group members in the minority, particularly in the KG treatment, this effect is not statistically significant.[15]

In sum, we find evidence of two types of gender-related bias in our known gender treatments. While we find no evidence of a general gender bias against women – women are ranked no worse than men, by neither self nor by others -- we do see more nuanced forms of gender biases. We find that gender stereotypes predict the likelihood of being chosen as the group representative. Individuals are more likely to be chosen in more gender congruent categories. This seems to stem mostly from stereotyping in ranking others and not from self-stereotyping. We also find that minority group members are less likely to be selected as the group representative. This seems to operate mostly through self-promotion: individuals rank themselves less favorably when they are in the gender minority.

[15] In an interacted model that explores whether the extent of these biases depends upon the gender of the ranker (i.e. individual *i*), we find no evidence that these effects vary significantly by ranker gender.

Table 4. The Determinants of the Probability of Group Member *i* Ranking Group Member *j* Ahead of the Other Group Member *k* Post-Group Interaction in the Chat Treatment

	All Groups			Mixed-Gender Groups		
Sample	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Gender j v k	4.956	5.021	4.869	4.690	6.362*	5.684
	(3.376)	(3.592)	(3.580)	(3.406)	(3.570)	(3.563)
Gender Congruence of Question j v k	12.09**	0.665	0.838	11.89**	0.162	0.580
	(5.166)	(5.708)	(5.617)	(5.205)	(5.628)	(5.590)
Own Gender Share in Group j v k	7.233	2.020	2.009	5.594	-0.646	0.540
	(6.744)	(7.174)	(7.081)	(6.788)	(7.188)	(7.104)
KG Treatment			-0.223			-2.442
			(3.647)			(4.216)
Gender j v k x KG			0.184			-0.891
			(4.880)			(4.892)
Gender Congruence j v k x KG			11.89			12.57*
			(7.536)			(7.498)
Own Gender Share j v k x KG			4.452			4.934
			(9.635)			(9.654)
Performance Controls	YES	YES	YES	YES	YES	YES
R-squared	0.0920	0.110	0.0727	0.125	0.138	0.100
Dep. Var. Mean	53.57	55.39	54.47	52.20	55.09	53.57
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Performance controls the differences in pre- and post-group individual scores of j relative to k. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

IV. B. The Importance of Free-Form Interaction

We have seen that knowing gender seems to be a necessary condition for generating gender biases. When gender is unknown, we find little evidence in support of any of our three hypotheses around bias. In this section, we ask whether the opportunity to interact is also a necessary condition for generating differences. In particular, we ask whether we would see similar patterns of representative selection if groups were not allowed to chat freely.

We design two control treatments to investigate this question. In each of these treatments, we eliminate the opportunity to chat, but keep other aspects of the design the same. In the Answer

Only treatment, we replace the group chat with an opportunity to view each of the other group member's pre-group stage individual answers. This allows us to ask whether, just seeing individuals' answers, provided with no justifications, expressed confidence, or advocacy, would be enough to generate gender biases. In the Answer + Confidence treatment, we transmit both the pre-group stage answers and the self-reported confidence of each group member. This allows us to ask whether answers, combined with a structured report of self-confidence, is enough to generate biases. In Appendix C, we re-run our main analysis on each of these control groups.

What we find is that neither Answers + Confidence nor Answers Only generates the patterns we see in the Chat treatment. We continue to find no evidence of preferential treatment of men relative to women in the control conditions. Furthermore, if individuals just view each other's answers, gender stereotypes and the gender composition of the group do not predict who is chosen as the group representative when gender is known. Similarly, even if individuals have a chance to view answers and self-reported confidence, gender stereotypes and group composition are still not significantly predictive of who is selected as the group representative. This suggests that there are features of the chat interaction itself that must contribute to our results. With this in mind, we turn our attention to better understanding those interactions.

V. ANALYSIS OF CONVERSATION DATA

Our experiment produced 276 natural language conversations between groups – a rich dataset that can yield new insights into the ways in which men and women communicate, advocate, and decide in groups. Perhaps most importantly, we can link these conversational patterns to incentivized decisions about who should be chosen as the group representative. We first attempt to understand behavior in these conversations, with a focus on detecting any gender differences. Then, we assess how behavior in these conversations impacts the likelihood of being chosen as the group representative.

V.A. GENDER DIFFERENCES IN OBJECTIVE CHAT CHARACTERISTICS

We begin with a brief overview of the trends in our conversation data. In particular, we code six objective variables: (1) number of engagements, measured as the number of times a participant enters anything into the chat interface in a given conversation; (2) volume of text, measured as the total number of characters typed in a given chat by a participant; (3) intensity of engagement,

calculated as the number of characters divided by the number of engagements; (4) share of other members convinced to submit an individual's own pre-group stage answer, which takes on the values of 0 (did not convince anyone else), ½ (convinced 1 other group member), or 2/2 (convinced both other group members); (5) an indicator for whether a given participant switched their answer in the post-group stage from their original pre-group submission; and (6) a count of the total number of new ideas (new relevant potential answers) submitted during a conversation by each group member. We also analyze more subjective assessments of the number of "weak" and "confident" words and expressions used in a given conversation (for example, "unsure" v. "definitely" – see Appendix D for details).

Subjects used the 60 second chat period to engage in lively and meaningful discussion. In no circumstances did we observe instances of abusive language, and in over 90 percent of the interactions, the chat submission was relevant to answering the question at hand (other than to start the interaction with a greeting). On average, subjects typed in 4 statements into the chat interface during a given conversation (i.e., average number of engagements was 4). Only 1 observation had no engagements, and the maximum number of engagements was 13 (see Appendix D1 for the distribution). We see only modest gender differences in objective conversation measures: while men have more engagements on average, women have a higher intensity of engagement on average, leading to quite similar numbers of total words used. Women are marginally more likely to use under-confident words, though this effect is only significant in the unknown gender treatment. Despite these similarities, men are more likely to suggest the answer that is ultimately submitted by the group when gender is known (49 v. 39 percent; t-test p-value of 0.059). Men are also able to convince more teammates (41 percent) to adopt their initial answer, while women are only able to convince 34 percent (though the gap is not significant; t-test p-value of 0.145). These differences are not present in the unknown gender treatment.

Appendix Table D2 asks how these factors impact the probability of being chosen as the group representative. We find that there are three important factors in predicting a favorable ranking in both treatments. First is the number of times a participant enters a statement into the chat interface (the level of engagement). More engaged participants are chosen more often as group representatives. Second is the ability to convince others to adopt one's initial answer. Note that these regressions condition on pre-group answer quality, so the fact that convincing others positively predicts being chosen is not simply picking up on some individuals having better pre-

group answers. Third, individuals who switch from their initial answer to a new answer in the post-group stage are significantly less likely to be chosen in both treatments. Being the one to suggest the ultimate group answer is also somewhat predictive of being chosen as group representative.

Table D3 in Appendix C also explores the heterogeneity of these effects by gender, gender congruence of question, and group gender composition. Overall, we see no large differences. It is clear that engagements, the share of others convinced, and not switching from your original pre-group answer positively predict being chosen as the group representative. Women who suggest the group answer and who use weak words are marginally less likely to be chosen in the KG treatment. The extent to which the rest of the conversation measures matter does not seem to vary much by gender, gender stereotype, or same gender share. Most crucially, the inclusion of these objective conversation measures does not explain any of the gender-related biases we observe. Even conditional on these objective measures, we find that individuals are significantly more likely to be chosen to represent the group in more gender congruent questions (see Table D2).

V.B. Methodology to Measure Subjective Chat Characteristics

Of course, these objective measures can only tell us so much about what happens during group chat. To give us more nuanced, subjective data about these conversations, we hired external coders to analyze the conversations. In a first pass analysis of the data, we asked approximately 1,000 workers from Amazon Mechanical Turk (AMT) to evaluate randomly chosen conversations, blinded to the participants' genders.[16] This coding experiment is reported in detail in Appendix E. The main findings of this initial experiment informed a more comprehensive second experiment that we report here, conducted to replicate and extend the findings of the first experiment.[17]

In Fall 2020, we recruited 500 Amazon Mechanical Turk (AMT) workers to read conversations and provide impressions (which we hereafter refer to as "coding"). Our main goal was to collect subjective perceptions of the members of each conversation, in terms of their

[16] Workers on AMT have been shown to exhibit similar behavioral patterns and pay attention to the instructions to the same extent as traditional subjects, particular in simple settings such as the one in our coding study (see Johnson and Ryan 2020 for the most recent evidence). Rand (2012) reviews replication studies that indicate that AMT data are reliable.

[17] We thank an anonymous referee who suggested this second experiment.

personalities, their conversational styles, and their contributions to the group's effort. Each AMT coder read five randomly-selected transcripts.

For each conversation shown to an AMT coder, we asked the coder to make a series of assessments of each member of the conversation. We focused on five features that were shown to capture meaningful variation in our first coding experiment: competent, assertive, warm, open-minded, and difficult to work with.[18] The 5-point scale ranged from "not at all" to "extremely" for each of these five dimensions. Note that four of these characteristics relate closely to the warmth-competence stereotype literature, that has found men to be stereotyped as assertive and competent, and women to be stereotyped as warmer and more open-minded (Fiske et al 2007). The main part of the survey was followed by a brief demographic questionnaire.

Because these are subjective views, we cannot incentivize coders according to the truth (i.e., we cannot induce an honest report about how assertive a group member is, given that there is no objective benchmark). So, to increase concentration and to motivate coders, we instead provide incentives through matching. Following participation, we matched each coder with another coder who faced one of the same chat transcripts. We then randomly selected one of the questions about that chat and compared the answers. If both coders gave the same answer to that question, the coder received an extra $1 in bonus payment, in addition to the $4 participation fee. In this way, we discourage coders from clicking randomly through the survey.

Coders were randomly assigned to one of two treatments. In the *blind* treatment, coders see all five of their conversations blinded to gender, with each member simply labeled Member 1, Member 2, or Member 3. In the *non-blind* treatment, coders see all five of their conversations with gendered pseudonyms attached to each member, corresponding to the actual self-reported gender of that conversation member.[19] Note that, within each treatment, conversations are randomly drawn for each coder from the entire population of conversations, independent of the original treatment for that conversation (known gender or unknown gender). This allows us to ask whether,

[18] In our first coding experiment, coders rated our participants on 17 dimensions. A factor analysis produced three factors: a factor that loaded heavily on competence and assertiveness, a factor that loaded heavily on warmth-related characteristics (warm, good-natured, tolerant), and a factor that loaded on stubborn, critical of others, and impeding the group's success. See Appendix E for more details.

[19] We use pseudonyms in our non-blind treatment to increase anonymity and to increase the likelihood of coders perceiving the correct gender. Following Manian and Sheth (2020), we use Social Security records to select 15 of the most popular baby girls' names and 15 of the most popular baby boys' names from 1990 – 1999, corresponding to the range of likely birth years for most of our college-aged laboratory conversationalists. These names are then randomly-assigned to conversation members, matching on gender, as conversations are randomly-drawn for a coder participating in the non-blind treatment.

holding fixed the sample of conversations, assessments of male and female members depend upon whether their gender is known.

V.C. GENDER STEREOTYPES IN CHAT DATA

Our question of interest is whether men and women vary in their communication styles, as rated by the coders. We consider assessments under the blind condition, and then ask, in an interacted model, whether these assessments look different when coders are not blind to gender. If men and women are assessed differently in our *blind* treatment (that is, when coders are unaware of conversation member gender), this would suggest true differences in how men and women behave in these conversations. By comparing assessments across our blind and non-blind treatments, we can ask whether perceptions of men and women are gender biased: does simply knowing someone's gender change how that individual's behavior is perceived? We also control for demographic information about our coders, round fixed effects (whether the conversation appeared first, second, third, fourth, or fifth for the coder), and conversation-level fixed effects.

Table 5 presents the results. When blind to the gender of the conversation member, our coders perceive women as significantly *more* assertive and competent than men, on average, in contrast with the classic stereotype. More in line with stereotypes, women are perceived as significantly warmer and directionally more open-minded than men, even when coders are blind to gender. There are no gender differences in how difficult men and women are perceived to be. If we consider blind evaluations to be good measures of true behavior in these interactions, we find that women in our study are, on average, more assertive, more competent, and warmer than men.

Next, note that men are rated very similarly on every dimension whether evaluated blind or non-blind (see the small and statistically insignificant coefficients on not-blind). However, perceptions of women along many of the dimensions are significantly different under the non-blind treatment as compared to the blind treatment. In particular, women are viewed as significantly less assertive and less competent when evaluated not blind compared to blind (see the coefficients on the interaction term female x not blind). These effects seem very much in line with gender stereotypes biasing the perceptions of women.

However, women are also perceived somewhat differently in the non-blind condition across the other three dimensions as well. Compared to blind assessments, women are also viewed as significantly less warm, directionally less open-minded, and directionally less difficult when

evaluated with gender information available. It seems to be the case that when coders know a woman's gender, their assessments of her are dampened toward the lower end of the scale for each of the dimensions. This effect does not occur for men. Keep in mind that across the two treatments, blind and not blind, conversations are randomly drawn from the same pool; thus, these differences can be attributed just to the labeling of conversation members with a gendered name.

Table 5: The Relationship between Observability of Gender and Perceived Communication Styles of Men and Women

	Assertive	Competent	Warm	Open-minded	Difficult
	(1)	(2)	(3)	(4)	(5)
Female	0.0622*	0.116***	0.110***	0.0411	-0.0168
	(0.0324)	(0.0336)	(0.0320)	(0.0330)	(0.0279)
Not Blind	0.0180	-0.0231	0.0136	-0.00842	0.0303
	(0.0549)	(0.0544)	(0.0594)	(0.0547)	(0.0616)
Female x Not Blind	-0.141***	-0.105**	-0.116**	-0.0624	-0.0594
	(0.0424)	(0.0455)	(0.0457)	(0.0481)	(0.0422)
R-squared	0.153	0.0935	0.106	0.104	0.249
Observations	7,403	7,403	7,403	7,403	7,403

Notes: Coefficients obtained using OLS for each of five communication factors coded as z-scores, independent of treatment. All specifications include fixed effects for group and conversation order; Rater demographic controls include gender, education, race/ethnicity, and whether they went to high-school in the United States. Robust standard errors clustered at the rater level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

We now turn our attention to how these conversationalist characteristics inform the selection of group representatives. In this section, we use the coder perceptions to address two questions. First, we ask which characteristics of individuals (assertiveness, competence, warmth, open-mindedness, difficulty) predict more favorable rankings of self and of others, conditional on answer quality. Second, we ask whether these characteristics help to explain the gender biases we documented in Section IV.

In Table 6, we predict how the self-rankings (Columns 1-4) and the relative rankings of others (Columns 5-8) depend upon the subjective characteristics of the conversationalists.[20] One important consideration is whether to use the blind or non-blind assessments from our coders. We match by treatment: using blinded assessments for the UG treatment and non-blinded assessments

[20] See Appendix D for this and subsequent analysis repeated for the total probability of being chosen as group representative as the outcome variable.

for the KG treatment. By doing so, we use the perceptions of conversation members that are more likely to match how they were actually perceived by fellow group members during the experiment. In all specifications, we control for true answer quality (the value of the answer she would submit as the group representative for self-ranking and the quality of answer of participant *j* would submit relative to *k* for relative rankings of others), and our standard demographic controls and fixed effects (paralleling Tables 3 and 4).

We first consider self-rankings. We see that in the KG treatment, none of the conversational characteristics predict the likelihood of a favorable self-ranking (Column 1). In the UG treatment, we see marginal evidence that more assertive and warmer group members rank themselves more favorably, while more open-minded members rank themselves less favorably (Column 2). Given these mostly null results, it is not surprising that these conversational characteristics do not explain the gender biases in self-rankings that we identified in Table 3. In particular, as we show in Column 3, even controlling for perceptions of assertiveness, competence, warmth, open-mindedness and difficulty, we still find a significant impact of own gender share on self-rankings in the KG treatment. If anything, the effect is stronger and more precisely estimated when we control for these perceptions. The fact that individuals are more likely to rank themselves favorably as the share of same gender group members grows does not seem to be correlated with behaving (at least as perceived by others) more assertively or more competently in those situations.

In Columns 5 – 8, we turn our attention to relative rankings of others. We find that perceptions of assertiveness and competence predict the likelihood of being ranked highly by others in both treatments (Columns 5-6), over and above true answer quality. In this way, stereotypically male-typed characteristics – assertiveness and competence – seem to be rewarded by groups. Warmth seems to have a marginally positive effect on rankings of others in the known gender treatment.

In Columns 7-8, we ask whether these characteristics explain the gender effects we documented in Table 4. That is, can the fact that individuals are more likely to be chosen in more gender congruent categories in the KG treatment be explained by, say, assertiveness, competence, or warmth varying with gender congruence? We find no evidence that this is the case, as the coefficient on gender congruence is unaffected by whether or not we control for these individual characteristics. Again, this suggests that the gender congruence effect is not well-explained by individuals behaving more assertively or more competently in more gender-congruent questions.

Table 6: The Effect of Communication Styles on the Probability of Favorable Self-Ranking (Panel 1) and on the Probability of Group Member *i* Ranking Group Member *j* Higher than the Other Group Member *k* (Panel 2) Post-Group Interaction in the Chat Treatment

	PANEL 1: Prob. of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)				PANEL 2: Prob. of Ranking j higher than k (Explanatory variables represent the difference btw j and k)			
	KG	UG	KG	UG	KG	UG	KG	UG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Assertive	-0.500	3.513*	-0.268	3.683**	6.455	8.094**	5.989	8.310**
	(2.134)	(1.791)	(2.129)	(1.813)	(3.954)	(4.037)	(3.934)	(4.067)
Competent	0.534	0.182	0.634	0.180	9.668**	5.979	9.121**	5.738
	(2.241)	(1.899)	(2.268)	(1.897)	(4.533)	(4.075)	(4.527)	(4.127)
Warm	-3.477	3.336*	-3.469	3.262*	7.565	-5.112	8.158*	-5.005
	(2.414)	(1.913)	(2.383)	(1.949)	(4.790)	(4.429)	(4.718)	(4.514)
Open	1.945	-3.091*	1.534	-2.994	-3.424	0.395	-2.680	0.697
	(2.406)	(1.866)	(2.430)	(1.883)	(4.843)	(4.153)	(4.914)	(4.158)
Difficult	0.814	-0.593	0.873	-0.404	3.588	-0.494	5.286	0.131
	(2.016)	(1.776)	(2.054)	(1.753)	(6.111)	(7.029)	(6.259)	(7.048)
Female			-1.528	0.889			3.810	5.175
			(1.900)	(1.882)			(3.545)	(3.649)
Gender Congruence of Quest.			2.783	-1.612			12.23**	0.153
			(2.770)	(2.834)			(5.261)	(5.859)
Own Gender Share in Group			7.665***	2.300			4.975	2.610
			(2.803)	(2.466)			(7.026)	(7.346)
Performance Controls	YES	YES	YES	YES	YES	YES	YES	YES
R-squared	0.109	0.123	0.130	0.126	0.100	0.133	0.118	0.138
Dep. Var. Mean	46.67	49.30	46.67	49.30	53.68	55.47	53.68	55.47
Observations	408	402	408	402	408	402	408	402

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for gender, age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

To summarize, the results of the coding experiment suggest that differences in participant behavior across conversations, at least as perceived by our coders, do not explain our main effects documented in Section IV. Conditional on interacting in very similar ways (as perceived by our coders) and conditional on the quality of answers they would submit as group representative,

individuals are still significantly more likely to be ranked more favorably by others in more gender congruent domains and more likely to self-promote when their own gender share in the group rises.

VI. THE ROLE OF BELIEFS

So far, we have documented that, conditional on answer quality, gender seems to play an important role in shaping who is selected to represent our groups. Individuals are more likely to be selected as group representative when the question is more gender congruent; this operates primarily through stereotyping of others. Individuals are also more likely to be chosen when they are in the gender majority; this is driven by individuals providing more favorable self-rankings as the share of their gender in the group increases. These results are not explained by differences in objective chat features (number of engagements, etc.), nor are they explained by individual communication styles or characteristics, such as perceived assertiveness, competence, and warmth.

In this final section, we attempt to understand whether these biases are well-explained by participant beliefs about answer quality, as described in our experimental framework. Or, are these distortions potentially driven by something other than beliefs, such as a taste for conforming with gender norms or a preference to reward members for their ideas or contributions in more gender congruent categories. While it will be impossible for us to completely pin down the relative roles of beliefs and preferences, we adopt a multi-pronged approach that begins to shed light on the possible channels.

In our main laboratory experiment, at the time of submitting their ranking of group members, each participant also submitted a belief on a 1-10 scale, indicating their confidence in each member to provide a high-scoring answer.[21] If beliefs of answer quality were the primary driver of our results, we expect that these assessments would have strong predictive power for rankings, and including them in the analysis would shrink the coefficients on the gender congruence of the question and same gender share. We test this hypothesis by returning to our main tables, and adding beliefs to the specifications.

Rather than considering the total probability of being chosen (which relies both on individual *i*'s ranking of herself and others' rankings of individual *i*), we split our analysis into

[21] Note that 10 indicates the highest possible confidence. These beliefs about others' ability to submit high-scoring answers were not incentivized. Note that the literature is mixed on whether complex elicitation mechanisms are more effective than simple introspection. See the discussion in Charness, Gneezy, and Rasocha (2020).

self-rankings – how individual *i* ranks herself, a la Table 3 – and rankings of others – how individual *i* ranks *j* relative to *k*, a la Table 4. We do so because for these specifications, there is a clear prediction for which beliefs should be the most relevant for the decision. For self-rankings, the beliefs of individual *i* about herself, about member *j,* and about member *k*, should determine her ranking. For rankings of others, the beliefs of individual *i* about member *j* and about member *k* should determine her relative ranking of the two.[22]

Columns 1 and 2 of Table 7 present the results for rankings of self. We replicate the specifications of Table 3, but add the three relevant belief measures: individual *i*'s belief in herself (indicated by Confidence in Member 1) and in the other two group members (indicated by Confidence in Member 2 and 3). We see that across both treatments, beliefs are strongly predictive of self-promotion decisions. As expected, individuals who report greater confidence in self provide more favorable self-rankings, while more optimistic beliefs about other group members predict less favorable self-rankings (note the large increase in r-squared relative to Table 3 as well). But, despite the significant predictive power of these beliefs, these belief measures have almost no impact on the coefficients related to gender biases. In particular, we continue to estimate a sizable and significant impact of being in the gender majority on self-promotion, even conditional on beliefs. Our data suggest that the gender composition of the group does not influence behavior through beliefs of answer quality. Individuals seem to have more of a taste for rewarding themselves with a more favorable self-ranking when they are in the majority.

We now turn our attention to the rankings of others. Columns 3 and 4 of Table 7 present the specifications of Table 4, but add the relevant belief measure: the difference between individual *i'*s belief in *j* and her belief in *k.* Once again, we estimate a large and significant impact of this belief on rankings of others across both treatments. We estimate that being one-point more confident in *j* relative to *k* increases the chances of ranking *j* above *k* by close to 10 percentage points. However, these beliefs explain only a small amount of the stereotyping effect in the KG treatment. Even controlling for beliefs, we find that an individual is significantly more likely to be ranked above another group member if the stereotype of the question benefits her relative to the other member. Thus, individuals seem to hold a preference for rewarding or recognizing group

[22] In contrast, for predicting total likelihood, one would need to account for all 9 belief measures: each member's assessment of herself and the other two members.

members in more gender congruent questions, above and beyond what can be explained by stereotyped beliefs about answer quality.

Table 7: The Effect of Beliefs about Performance on the Probability of Favorable Self-Ranking and on the Probability of Group Member *i* Ranking Group Member *j* Higher than the Other Group Member *k* Post-Group Interaction in the Chat Treatment

	PANEL 1: Prob. of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)		PANEL 2: Prob. of Ranking j higher than k (Explanatory variables represent the difference btw j and k)	
	KG	UG	KG	UG
	(1)	(2)	(3)	(4)
Female	-1.424	1.438	2.529	5.375*
	(1.545)	(1.584)	(2.833)	(3.032)
Gender Congruence of Question	2.369	-1.773	9.162**	2.502
	(2.315)	(2.335)	(4.344)	(4.800)
Own Gender Share in Group	6.957***	-0.563	8.825	-3.498
	(2.386)	(1.969)	(5.694)	(6.046)
Confidence in Member 1	4.706***	4.845***		
	(0.396)	(0.451)		
Confidence in Member 2	-1.657***	-1.323***		
	(0.321)	(0.294)		
Confidence in Member 3	-1.420***	-1.211***		
	(0.384)	(0.324)		
Confidence in Member j v. k			9.680***	9.175***
			(0.607)	(0.632)
Performance Controls	YES	YES	YES	YES
R-squared	0.362	0.366	0.340	0.356
Dep. Var. Mean	46.67	49.26	53.57	55.39
Observations	408	402	408	402

Notes: Coefficients obtained using a linear probability model. All specifications include fixed effects for round and part; demographic controls for gender, age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

This surprising set of results raises an important question: if biased beliefs of answer quality are not driving the gender biases we observe, what is? The first, less interesting explanation is that it may be that beliefs are simply too noisily estimated in our experiment; this type of

measurement error could lead us to underestimate the role of beliefs.[23] But, given that beliefs do seem to have significant predictive power, but have little impact on our gender coefficients, we expect that this cannot be the primary explanation. The second explanation is that rankings are reflecting other substantive considerations, such as a desire to reward or recognize certain group members for their contributions, or a preference for certain types of people being chosen to represent the group.

In order to investigate the second possibility, we asked the third-party coders in the follow-up experiment reported in Section V to also assess other beliefs and preferences that may be relevant in our framework. In particular, after a coder completed the assessments of each group member on assertiveness, etc. for a given conversation, we asked them three simple questions:

1. Which group member made the most valuable contributions to the group?
2. If you could reward one group member, giving them extra compensation, who would you choose?
3. If you had to choose one group member to answer this question on your behalf, where your compensation would depend on the answer they would submit, who would you choose?

We think of these questions as getting at the three primary motives that could drive rankings: a desire to recognize someone who made valuable contributions to the group, a preference to reward a given group member more generally, and finally, a belief that a given group member is likely to submit a good answer. This third measure is in line with the beliefs we capture in our original experiment, allowing us to ask, in the second setup, with potentially less measurement error, whether beliefs have any additional predictive power.

We start by simply documenting how the coders answered these questions, across both the blind and not blind versions of the coding experiment. When coders are blind to gender, female group members are 5 percentage points more likely than male group members to be singled out as having made valuable contributions to the group, and 4 percentage points more likely to be chosen to (hypothetically) receive extra compensation. There are no significant gender differences in

[23] There is an increasing focus on understanding the role of beliefs in driving gender differences, for instance in the realm of competitiveness (Charness, Rustichini, and van de Ven 2018; Gillen, Snowberg, and Yariv 2019; van Veldhuizen 2017). In past work on willingness to volunteer ideas, beliefs have also been shown to play a critical role (Coffman 2014).

being chosen to have their answer counted (i.e., betting on a given group member to answer the question on the coder's behalf). Similar to what we found with assessments of personal characteristics, knowing gender seems to color coders' decisions. In particular, male group members are significantly more likely to be identified as having made valuable contributions to the group when the coders know gender as compared to when they do not. Women, on the other hand, have a lower likelihood of being recognized as having made valuable contributions when coders are aware of gender. This suggests gender bias, among our coders, in terms of how contributions are perceived.

Table 8: The Relationship between Observability of Gender and Recognition by Others

	Made Valuable Contributions	Deserves Extra Compensation	Would Bet of Their Answer
	(1)	(2)	(3)
Female	0.0532***	0.0410**	0.0134
	(0.0206)	(0.0190)	(0.0201)
Not Blind	0.0343***	0.0141	0.0130
	(0.0119)	(0.0114)	(0.0115)
Female x Not Blind	-0.0673***	-0.0276	-0.0255
	(0.0233)	(0.0224)	(0.0225)
R-squared	0.0020	0.0010	0.0004
Observations	7,403	7,403	7,403

Notes: Coefficients obtained using OLS. Dependent variable mean is 0.33. All specifications include fixed effects for group and conversation; Rater demographic controls include gender, education, race/ethnicity, and whether they went to high-school in the United States. Robust standard errors clustered at the rater level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Does this help to explain the gender-related biases in who is selected to represent the group in our main experiment? To address this question, we ask whether the coders' perceptions of participants' "deservingness" along these different dimensions predict the probability of being chosen as the group representative. Following the approach of Table 6, the analysis in Table 9 replicates our main results from Tables 3 and 4, but adding the average of coder assessments of the group member for each of the three questions to the model.[24] Again, we match on treatment, using blind assessments for the UG treatment and non-blind assessments in the KG treatment, in order to add to the model the coder perceptions that are most likely to match the perceptions of the

[24] We document the overall effect on the probability of being chosen as group representative in Appendix D.3.

group members themselves, hopefully maximizing explanatory power. While it is not clear that these perceptions should predict self-rankings (that is, that players perceived to have made valuable contributions or who were perceived to deserve extra compensation would also rank themselves higher), it seems likely that these should predict how they are ranked by others.

Table 9: The Effect of Third-Party Recognition on the Probability of Favorable Self-Ranking and on the Probability of Group Member *i* Ranking Group Member *j* Higher than the Other Group Member *k* Post-Group Interaction in the Chat Treatment

	PANEL 1: Prob. of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)				PANEL 2: Prob. of Ranking j higher than k (Explanatory variables represent the difference btw j and k)			
	KG	UG	KG	UG	KG	UG	KG	UG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Made Valuable Contributions	5.382	6.109*	4.753	6.194*	6.389	8.936	6.663	8.677
	(4.867)	(3.657)	(4.883)	(3.736)	(6.999)	(6.313)	(7.005)	(6.360)
Deserves Extra Compensation	9.275**	-0.0182	9.261**	0.188	0.514	-3.503	-0.195	-3.498
	(4.652)	(4.099)	(4.667)	(4.106)	(7.332)	(6.955)	(7.281)	(7.041)
Bet on Being Chosen	3.488	6.416*	3.465	6.258*	18.12***	8.251	17.11***	8.554
	(4.223)	(3.318)	(4.205)	(3.342)	(6.577)	(5.852)	(6.540)	(5.991)
Female			-1.568	1.167			4.669	4.758
			(1.863)	(1.844)			(3.472)	(3.589)
Gender Congruence of Quest.			2.271	-1.190			10.24**	1.853
			(2.678)	(2.812)			(5.176)	(5.781)
Own Gender Share in Group			7.113**	1.936			4.554	2.854
			(2.752)	(2.436)			(6.825)	(7.233)
Performance Controls	YES	YES	YES	YES	YES	YES	YES	YES
R-squared	0.142	0.129	0.160	0.132	0.112	0.122	0.127	0.127
Dep. Var. Mean	46.67	49.30	46.67	49.30	53.68	55.47	53.68	55.47
Observations	408	402	408	402	408	402	408	402

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for gender, age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

The impact of the three measures of deservingness on rankings does not follow a consistent pattern. When it comes to self-rankings (Columns 1 and 2 of Table 9), each of the three dimensions

– being perceived as having made valuable contributions, being chosen to receive extra compensation, and being chosen to have your answer submitted by the coders – seems to at least weakly predict a favorable self-ranking. But, the effects vary substantially across the treatments. More centrally, Columns 3 and 4 document that none of these perceptions explain the gender bias we observe in self-rankings: even conditional on these perceptions, individuals rank themselves more favorably when the share of same gender group members is larger.

Turning to the rankings of others, we see that being perceived as having made relatively more valuable contributions is associated with being ranked above another group member, but the effect is not significant. In the KG treatment (Column 5), we see a large and significant positive impact of the coders' choosing to submit that group member's answer on her being more likely to be ranked ahead of another group member, but the effect is not present in the UG treatment.

Most centrally, Column 7 shows that gender biases remain unexplained by these new measures. It continues to be the case that individuals are ranked more highly when the question is more gender congruent. This seems to operate on top of any stated beliefs about who is likely to submit the best answer, who is perceived to have made particularly valuable contributions to the group, or who is perceived to be deserving of extra compensation. We can only speculate as to what may contribute to these patterns. It may be that individuals have a (perhaps implicit) taste for choosing more gender congruent representatives, or are conforming to a perceived norm about who the most appropriate representative might be. These interesting questions merit consideration in future work.

VII. DISCUSSION

Our paper explores the ways in which gender-related biases shape group decision-making. In our framework, we allow for free-form chat across group members, providing additional insights into how biases operate. We find that individuals are less likely to be rewarded for their ideas in gender incongruent domains when gender is known. This is a result of discrimination by fellow group members, whose relative rankings of their fellow group members depend heavily on gender stereotypes. We also observe differences in the propensity to self-promote in the known gender treatment: individuals are more likely to rank themselves favorably when they are in the gender majority.

We use a wealth of data and a variety of approaches to try to unpack the drivers of these results. Neither stereotyping of others nor diminished self-promotion when in the minority are well-explained by true differences in answer quality, or by beliefs about answer quality. Furthermore, neither objectively nor subjectively coded features of communication styles explain these patterns. Further investigation should consider what types of preferences might be relevant in explaining these types of persistent gender biases.

A natural question to ask is what the results imply for earnings in our experiment. First, we point out that, relative to our control treatments, groups earn significantly more in our chat treatments, suggesting that there is (perhaps not surprisingly) substantial value in deliberation. Second, when we compare across the known and unknown gender chat treatments, we find no significant differences in group earnings. Thus, while we find that stereotypes and group composition influence representative selection, over and above true answer quality, this does not have a significant negative impact on efficiency. Beyond efficiency, our results have important implications for the distribution of earnings within our groups. Because of the financial reward for serving as the group representative, individuals in the known gender treatment have higher earnings in the more gender congruent domains, and when they are in the gender majority rather than in the minority.

In many ways, our environment comes closer to "real world" settings than past experimental work in this space, allowing for free form communication in a subjective decision-making problem. The fact that we find distortions in contribution and recognition in this environment raises important questions about how these forces might fuel gender differences in workplace outcomes. Our work suggests a need for structuring group decision-making in a way that assures that the most talented members both volunteer and are recognized for their contributions, despite gender stereotypes.

REFERENCES

Alan S., Ertac S., Kubilay E., Loranth, G. 2017. "Understanding Gender Differences in Leadership." Working paper.

Born, A., Ranehill, E., Sandberg, A. 2020. "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" *Review of Economics and Statistics:* 1 – 46.

Bertrand M, Goldin C, Katz LF. 2010. "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics*, 2 (3): 228-255.

Brandts, J., Charness, G., Ellman, M. 2015. "Let's Talk: How Communication Affects Contract Design," *Journal of the European Economic Association*, 14 (4): 943-974.

Bordalo, P., Coffman, K. B., Gennaioli N., Schleifer, A. 2019. "Beliefs about Gender," *American Economic Review* 109(3): 739-773.

Bursztyn, L, Fujiwara T, Pallais A. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, 107 (11): 3288-3319.

Catalyst. 2020. "Pyramid: Women in S&P 500 Companies," http://www.catalyst.org.

Charness, G., Rustichini, A. 2011. "Gender Differences in Cooperation with Group Membership," *Games and Economic Behavior*, 72 (1): 77-85.

Charness, G., Rustichini, A., and van de Ven, J. 2018. "Self-confidence and strategic behavior," *Experimental Economics*, 21 (1, no. 4), 72–98.

Charness, G., Dugwenberg, M. 2010. "Bare Promises: An Experiment." *Economics Letters*, 107 (2): 281 – 283.

Charness, G., Gneezy, U., and Rasocha, V. 2020. "Experimental methods: Eliciting beliefs," forthcoming in *Journal of Economic Behavior and Organization*.

Charness, G. Rustichini, A. and van de Ven, J. 2018. "Self-confidence and Strategic Behavior." *Experimental Economics*, 21 (1, no. 4), 72–98.

Chen, J., and Houser, D. 2019. "When are women willing to lead? The effect of team gender composition and gendered tasks," *The Leadership Quarterly* 30(6): 101340.

Coffman, K. B. 2014. "Evidence on Self-stereotyping and the Contribution of Ideas," *The Quarterly Journal of Economics*, 129(4): 1625–1660.

Dreber, A., von Essen, E., Ranehill, E. 2014. "Gender and Competition in Adolescence: Task Matters," *Experimental Economics* 17 (1): 154–72.

Dugar, S., Shahriar, Q. 2018. "Restricted and Free-form Cheap-talk and the Scope for Efficient Coordination," *Games and Economic Behavior*, 109 (C): 294—310.

Gillen, B., Snowberg, E., and Yariv, L. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy*, *127*(4), 1826–1863.

Grosse, N. D., Riener, G., Dertwinkel-Kalt, M. 2014. "Explaining Gender Differences in Competitiveness: Testing a Theory on Gender-Task Stereotypes," Mimeo, 1–35.

Goldin, C., Kerr, S. P., Olivetti, C., Barth, E. 2017. "The Expanding Gender Earnings Gap: Evidence from the LEHD-2000 Census," *American Economic Review: Papers and Proceedings,* 107 (5): 110-114.

Hernandez-Arenaz, I. 2018. "Stereotypes and Tournament Self-Selection: A Theoretical and Experimental Approach," University of the Balearic Islands Working Paper.

Isaksson, S. 2018. "It Takes Two: Gender Differences in Group Work," Working paper.

Johnson, D. and Ryan, J. B. 2020. "Amazon Mechanical Turk workers can provide consistent and economically meaningful data," *Southern Economic Journal*, 87 (1), 369–385.

Michelmore, K., Sassler, S. 2016. "Explaining the Gender Wage Gap in STEM: Does Field Sex Composition Matter?" *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(4): 194–215.

Niederle, M., Vesterlund, L. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *The Quarterly Journal of Economics* 122 (3): 1067–1101.

Niederle, M. 2016. "Gender," in *The Handbook of Experimental Economics 2*, Kagel John, Roth Alvin E., eds. (Princeton, NJ: Princeton University Press, 2016).

Rand D.G. 2012. "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments," *Journal of Theoretical Biology*, 299: 172–179.

Sarsons, H. 2017. "Recognition for Group Work: Gender Differences in Academia," *American Economic Review: Papers and Proceedings,* 107 (5): 141-145.

Shan, X. 2020. "Does Minority Status Drive Women Out of Male-Dominated Fields?" *Working paper.*

Shurchkov, O. 2012. "Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints," *Journal of the European Economic Association* 10 (5): 1189–1213.

Shurchkov, O., Eckel C. C. 2018. "Gender Differences in Behavioral Traits and Labor Market Outcomes," in *The Oxford Handbook of Women and the Economy*, Averett Susan L., Argys Laura M., Hoffman Saul D., eds. (Oxford, UK: Oxford University Press, 2018).

Shurchkov, O., van Geen, A. 2019. "Why Female Decision-Makers Shy Away from Promoting Competition," *Kyklos*, 72 (2): 297-331.

van Veldhuizen, R. 2017. "Gender Differences in Tournament Choices: Risk Preferences, Overconfidence or Competitiveness?", Rationality and Competition Discussion Paper Series 14, CRC TRR 190 Rationality and Competition.