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

GENDER STEREOTYPES IN DELIBERATION AND TEAM DECISIONS[*]

Katherine Coffman[†] Clio Bryant Flikkema[‡] Olga Shurchkov[§]

Harvard Business School Wellesley College

First Draft: November 2018

Current Draft: January 2019

Abstract:

We run an experiment that features a novel task with deliberation to explore how stereotypes shape group decision-making. We find that women are less likely to be rewarded for their ideas in male-typed domains when gender is known. This is partly due to discrimination, and partly due to differences in self-promotion. External analysis of the chat data provides further insights. Though men and women do not vary in their communication styles, coders display pervasive stereotypes, associating warmth with women and competence and negativity with men. We also find that warmer participants, particularly warmer women, are under-rewarded by their groups.

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

JEL Classifications: C90, J16, J71

[*] We thank Manuela Collis, Kyra Frye, and Marema Gaye for excellent research assistance. We also wish to thank Kristin Butcher, G. Kartini Shastry, and the participants of the work-in-progress seminar at the University of Massachusetts Amherst and the Wellesley College Economics Research Seminar for helpful comments. Flikkema and Shurchkov gratefully acknowledge financial support from Wellesley College Faculty Awards. All remaining errors are our own.

[†] Corresponding author. Negotiations, Organizations, & Markets Unit, Harvard Business School, Soldiers Field Boston, MA, USA, katiebcoffman@gmail.com

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

[§] 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 contexts. In this paper, we explore gender differences in the ways in which men and women communicate in team decision-making problems. We ask whether there are differences in the propensity of men and women to self-promote themselves and their ideas in these contexts, and whether they are equally likely to be recognized and rewarded for their ideas.

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 2018). The gender gap in representation as well as 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 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 (2018) and Chen and Houser (2017) 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 reward ideas. We explore these questions by designing a controlled laboratory experiment that utilizes free form chat among group members. In our environment, teams brainstorm answers to questions that vary according to the gender perception 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, admitting 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. We compare behavior across two free form chat treatments that vary whether gender is revealed to fellow group members. This allows us to cleanly measure discrimination in how contributions are valued.

We find that even though there are no gender differences in individual ability to answer the questions, gender stereotypes play a significant role in which ideas are rewarded in the known-gender treatment. As the maleness of the question increases, women are significantly less likely to be selected to answer on behalf of the group (conditional on the quality of their contributions). They are also significantly less likely than men to self-promote in the known-gender treatment, particularly when they are the lone woman in the group. In comparison, there are no gender differences when gender is unknown.

Our second main contribution comes from the analysis of the natural language conversation data to provide further insights into the team decision-making process. Third-party external evaluators read and rate the contributions of each group member, blinded to the gender of the participants. Interestingly, and perhaps contrary to widely-held beliefs, we find no significant gender differences in the way in which men and women communicate. Despite this, we find a powerful role for gender stereotypes in the raters' perceptions: evaluators of conversations are significantly more likely to believe that a warm participant is female, and that a negative or critical participant is male. Male raters also believe that members who are judged as competent are significantly less likely to be female. We explore the returns to warmth, competency, and negativity in our group decision-making task and find that even though warmth is a strong predictor of being a good group representative, warmer participants are less likely to be promoted to group representative. This is particularly true for warm women when gender is known.

Our results are consistent with a growing literature showing the importance of stereotypes for economic outcomes. For example, Shurchkov (2012), Dreber et al. (2014), and Grosse et al. (2014) show that gender gaps in willingness to compete become substantially smaller and insignificant in the context of a more female-typed task as compared to a stereotypically male-typed task used by Niederle and Vesterlund (2007). Similarly, Iñigo 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 to subsequent evaluators (Shurchkov and van Geen 2018). Public observability in the presence of gender stereotypes has also been shown to significantly decrease women's willingness to lead (Alan et al 2017), willingness to compete (Buser et al 2017), and to express ambition (Bursztyn et al. 2017). Our work suggests that willingness to self-promote also depends upon the observability of gender.

II. THE EXPERIMENT

II.A. THE TASK

Participants in our experiment play multiple rounds of a *Family Feud* style task.[1] To our knowledge our study is the first to use a modified *Family Feud* game in an economic experiment. The task was chosen to mirror the real-world properties of group discussion settings. In this task, the best solution can be determined by logical reasoning, but there is room for disagreement among the participants because some reasonable ideas are better than others, and points are based upon subjective beliefs.

In particular, the goal of the game is to guess an answer to the question that would be frequently given by others. Specifically, the Family Feud questions we source have been previously shown to a 100-person survey panel, who each gave answers to the question. These 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 participants 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. Consider the example below:

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*

[1] 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

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 the team some points, albeit less than the top answer. Only answers that received two or more survey responses could count for points. If the answer submitted did not appear in the table of answers, the subject received zero points.

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 in Appendix G). AMT participants provided answers to a subset of 20 possible Family Feud questions. And, they provided their perception of the gender stereotype for each question, indicating for each question on a -1 to 1 scale whether they believed men or women would be better at answering that particular question. Using this data, we selected 8 questions that are perceived as gender stereotyped (four female-typed and four male-typed). These were subsequently randomly assigned at the session level, one for each round of the experiment. The extent to which a question is perceived to favor men relative to women is one of our main variables in subsequent analysis, coded as “*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).

II.B. Experimental Design

Each session of the experiment consisted of two parts, each containing four rounds of interactions, using one of eight *Family Feud* questions. In each round, participants were randomly re-matched in groups of three, using stranger matching. All interaction took place via private computer terminals.

Our primary treatment variation is whether or not gender information is made available to participants. In the unknown-gender treatment, participants were identified in each round by a randomly-generated ID number. In the known-gender treatment, 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 treatment. They were encouraged to use their real name, but participants were able to select any name they wished.[2] This name was then used throughout the part to identify them to their fellow group members. Second, we did a verbal roll call, in which groups were

[2] 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.

announced out loud, and each member of the group was called by 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 2018).

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. Chat entries are identified either by names in (known-gender) or by ID number (unknown-gender).

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 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 the 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).

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 will receive 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 positions of leadership or recognition 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.

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 individual answer in pre-group stage, an 80% chance they were paid for the group answer given by the selected representative, and a 10% chance they were paid for their individual answer in the post-group stage. In addition, the person selected as the "group representative" received a bonus payment of $2, providing a material incentive to be chosen.

In addition to our main chat treatments, we have two control treatments aimed at understanding the mechanisms at work. We describe these control treatments in detail in Appendix D. In each session, subjects participated in exactly two treatments, one in each part. In every case, one of these treatments was a Known Gender treatment and the other was an Unknown Gender treatment. In total 207 subjects participated in our main chat treatments, 105 in the Known Gender version and 102 in the Unknown Gender version. Data was collected at the CLER laboratory at Harvard Business School between September 2017 – May 2018.

III. RESULTS

In the Appendix, we provide summary statistics (Appendix A) and analysis of the pre-group stage (Appendix B). We find no gender differences in pre-group stage ability in our data: men and women submit equally high-scoring answers on average, regardless of the gender-type of the question. On average, there are no gender differences in beliefs of own ability, as captured by the pre-group stage confidence question, conditional on measured ability. But, our estimates suggest that women respond differently to maleness in the known-gender treatment compared to the unknown-gender treatment. While women are estimated to grow directionally *more* confident as maleness increases when gender is unknown, they are estimated to become significantly less confident as maleness increases when gender is known. It seems that the salience of gender in the known-gender treatment may encourage stereotypical thinking in terms of beliefs about own ability among women.

The ranking provided by group members is our key outcome variable from the group decision-making task. We are interested in whether there is a gender difference in the propensity to be selected to represent the group, and if so, what drives this difference. Is representative selection a function of gender stereotypes? And, to what extent do differences in self-promotion (how an individual ranks oneself) versus recognition by others (how an individual is ranked by her peers) contribute to the gender gap? How does this depend upon the observability of gender?

Our main question is whether there are gender differences in the probability of being chosen as the group representative. Table 1 explores the determinants of this probability, calculated as the average probability of being chosen given the rankings of each group member. We control for an individual's pre-group stage ability, as measured by the quality of her individual pre-group answer, the quality of her individual 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 answer versus average points earned by all answers in the group), and the quality of her individual answer relative to the maximum quality of individual pre-group answers in her group (all measured in terms of points earned). We also control for part and round fixed effects, as well as demographic characteristics, and we cluster standard errors at the group level.

We find that women are not significantly less likely to be chosen on average overall (Column 1). But, the gender gap in the probability of being chosen is impacted by the maleness of the question in the known-gender treatment (Column 2). We zoom in on the known-gender treatment in Columns 3-4, documenting the existence of stereotyping. Men are more likely to be chosen as the group representative as the maleness of the question increases, while women are directionally less likely to be chosen as maleness increases. Furthermore, a given woman is more likely to be chosen as the share of other women in her group increases (Column 4).

[TABLE 1 ABOUT HERE]

This probability of being chosen is shaped by two distinct factors: the participant's self-ranking (her propensity to self-promote) and the ranking she receives from others. In the Appendix, we decompose these two channels. We find that there are no gender differences in the propensity to self-promote on average. However, we see that group composition seems to shape self-promotion behavior in the known-gender treatment. In groups with no other women, we estimate that women give themselves a 7.5 pp lower probability of answering for the group compared to men. But, as the share of other women in the group increases, the gender gap is reversed. Both men and women self-promote less often when they are the minority group member than when they are in the majority (see Table C1).

In terms of ranking by others, we find no significant differences in how men and women are ranked in the unknown-gender treatment. But, in the known-gender treatment, women are significantly less likely than men to be chosen as the maleness of the question increases, suggestive of stereotyping (see Table C2). This combination of differences in self-promotion behavior and discrimination from others leads to the gender gap in serving as group representative in the known gender treatment.

We conducted two control treatments to further understand the contributors to gender gaps in group representation. In these treatments, we shut down the chat channel and restrict the nature of interaction among group members, either by only displaying pre-group stage answers to each group member or by displaying pre-group stage answers and self-confidence assessments to each group member. In this way, we can better understand whether gender gaps would be as likely to emerge absent the deliberation stage, conditional only on ability and self-reported confidence. We present the results in Appendix D, where we show that essentially no gender gaps emerge absent a deliberation stage. Thus, the conversations seem to be crucial in driving the gender gaps we observe.

IV. ANALYSIS OF CONVERSATION DATA

IVA. Methodology

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. To make sense of this data in an objective and tractable way, we recruited 1000 Amazon Mechanical Turk (AMT) workers to read these conversations and provide impressions of the contributions made.[3]

[3] Workers on AMT have been shown to exhibit similar behavioral patterns and pay attention to the instructions to the same extent as traditional subjects (Paolacci et al. 2010; Germine et al. 2012). Rand (2012) reviews replication studies that indicate that AMT data are reliable. We used randomly placed attention checking questions in order to ensure full attention. The final dataset contains valid responses of 985 AMT raters.

Each AMT participant read three randomly-selected transcripts. Importantly, within each conversation, members were labeled simply as Member 1, 2, or 3. That is, we blind AMT participants to gender.

For each conversation shown to the participant, she was asked a series of questions about each member of the conversation, both communication-style focused and performance focused (see Appendix for instructions). Following the warmth-competence literature (Fiske et al 2007), we asked participants to evaluate members on three dimensions of warmth (warm, tolerant, good-natured) and competence (competent, intelligent, confident). We also asked about how assertive and passive the member was, whether they were supportive or critical of others, and how stubborn they seemed. These 11 personality traits were presented in one block for each member, in an order randomized at the individual level.

We also asked AMT participants performance-oriented questions: to what extent each group member contributed to group success, did a good job voicing their ideas, advocated to be chosen by the group, impeded the group's success, advocated for their preferred answer, and had their ideas listened to by the group. These were again organized into one block and randomized at the individual level.

The 5-point scale ranged from "not at all" to "extremely" for all questions. At the end of each conversation question set, the AMT participants had to choose which of the three members they would vote as the "MVP (most valuable player)". Finally, after the last of the three conversations they rated, the AMT participants guessed the gender of each of the members in that chat. We only asked this question once at the very end of the survey to not give away our interest in gender.

Following participation, we matched each participant with another participant 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 participants gave the same answer to that question, the participant received an extra $1.50 in bonus payment, in addition to the $2 participation fee.

In order to categorize questions into broader explanatory factors that are orthogonal to one another, we performed a principle component decomposition. It yielded three factors. Factor 1 loads heavily on competency, confidence, and assertiveness – aligning closely with the competence dimension identified by Fiske et al (2007); Factor 2 on warmth, good-naturedness, being supportive of others, and tolerance – aligning with the warmth dimension identified by Fiske et al (2007); and Factor 3 on the more negative traits of stubbornness, being critical of others, and impeding success. Note that the components are z-scores.(see Appendix Table E1 for details).

IVB. Gender Stereotypes in Chat Data

Our first step is to document whether men and women communicate differently in our conversations. In Table 2, we show that men and women are rated as identically competent, warm, and negative based upon their conversation contributions. That is, when blind to gender, coders perceive men and women as the same on all three dimensions on average. Note that these results are unchanged if we consider only the known gender treatment (or unknown gender treatment).

[TABLE 2 ABOUT HERE]

Recall that we ask our coders to guess the gender of the members of the conversations. Thus, we can ask what predicts the probability that a coder believes that a member is female. Table 3 reports the estimates from an OLS regression that predicts the likelihood that an AMT rater guessed a given participant was female from their evaluation of that member in terms of the three conversation factors we identified – Competence, Warmth, and Negativity. Importantly, these estimates are not causal: we cannot rule out that an unmeasured factor or conversation feature leads the coder to both evaluate the member in a particular way and guess that he or she is female. These estimates simply tell us which factors are correlated with a coder believing someone is female.

Column 1 shows that members viewed as warm (coded in Factor 2) by the rater are more likely to be believed to be female, while members viewed as negative (coded in Factor 3) are more likely to be believed to be male. Thus, while there are no *actual* differences in how men and women seem to communicate in these settings (at least as perceived by our coders), the coders hold strong stereotypes about the behavior that is more typical of men or women. Being warm is strongly associated with being perceived as female; being critical is strongly associated with being perceived as male.

We can also disaggregate the analysis by rater gender. Here, we see that these stereotypes regarding warmth and negativity are exhibited by both male and female raters. Interestingly, we see that for male raters, viewing a member as competent is associated with a significantly lower probability that the rater believes that member is female. That is, male raters (falsely) associate being more component with being male. This is not true for female raters. Given the inaccuracy of these stereotypes, it is perhaps not surprising that the raters are on average quite bad at correctly guessing gender: less than 45% of women are correctly identified as women.

Summing up the evidence on gender stereotypes, we see that raters provide nearly identical ratings of men and women in our data on competence, warmth, and negativity. Yet, when asked to guess gender, the same coders incorrectly believe that those individuals that they rated as warmer or less negative (and in the case of male coders, less competent) are more likely to be women.

[TABLE 3 ABOUT HERE]

We now turn our attention to how the chats inform the selection of group representatives. The group does best by selecting the individual who will submit the highest scoring answer in the post-group stage. In our data, we observe the post-group stage answer from each individual in the group, allowing us to judge whom would be the best choice within each group. In this section, we use this valuable data to address two questions: first, which chat factors predict talent as a group representative? and, second, which chat factors do groups actually seem to rely on selecting their group representative?

To address the first question, we regress the quality of each individual's post-group stage answer – the answer they would submit for the group if chosen as the representative – on the average factor ratings provided for that person in that conversation by our external coders, including our standard set of demographic controls and fixed effects. Note that here we omit pre-group stage ability measures, due to their correlation with the competence dimension.[4] Table 4 presents the results. In Column 1, we observe that directionally, all three factors are positively predictive of submitting a higher scoring answer. The strongest factor is warmth: warmer participants are significantly better group representatives ($p<0.004$). Competence is also a significant predictor ($p<0.05$), but the effect size is roughly 40% smaller.[5] These findings are similar across the two treatments and across men and women, with warmth consistently being the

[4] See Appendix Table F1, for specifications that include pre-group stage ability measures.

[5] Note that controls for distribution of pre-group performance absorb the positive effect of competence on post-group performance. This can likely be explained by the correlation between ability, as measured by the individual answer in pre-group stage and competence. The positive effect of warmth, on the other hand, is robust to the inclusion of pre-group performance controls (see Appendix Table F1).

directionally the largest predictor of being a good group representative, though it has the largest effect in the known gender treatment, and a marginally larger effect for men than women (see Columns 3 – 6). Thus, our results suggest that, if groups are attempting to maximize group earnings, they should be selecting group representatives who are competent, but even more importantly warm.

[TABLE 4 ABOUT HERE]

With this in mind, we return to the rankings provided by the group members and ask how the three factors of competence, warmth, and negativity predict how members were ranked in the experiment *by their fellow group members*. That is, consistent with Table 4, do groups do a good job of selecting warm, competent group representatives? We explore this in Table 5, predicting a participant's rating by another group member from her average rating on each factor for that conversation. We include the same set of controls as Table 4, to keep the analysis as parallel as possible. First, we pool across the two treatments in Columns 1 and 2. We see that competence is strongly predictive of receiving a favorable ranking from others. This is true in both the KG and UG treatments (Columns 3—6). Thus, groups seem to reward competence. However, they do not reward warmth. In fact, warmth has a *negative* impact on ranking in the KG treatment (Columns 3 and 4), and an insignificant negative effect in the UG treatment (Columns 5 and 6). Interestingly, Column 4 reveals that only women in the KG treatment are penalized for warmth. On the other hand, men in the KG treatment (Column 4) and men and women in the UG treatment (Column 6) receive no such penalty. Negativity has a directionally negative impact on ranking in both treatments, with women being directionally more penalized but not significantly so. These results are robust to the inclusion of controls for distribution of pre-group performance. Furthermore, controlling for our outcome measure of Table 4 – actual talent as the group representative, i.e. post-group stage individual performance – does not affect the results in Table 5 and, importantly, has a negligible effect on the probability of being chosen (both in significance and magnitude, see Appendix Table F2). This once again points to the fact that groups are not particularly capable of selecting good group representatives.

[TABLE 5 ABOUT HERE]

In sum, groups seem to being getting it wrong in a predictable way. While the data tells us that warmth is strongly associated with being a high scoring group representative, groups are, if anything, less likely to select warm individuals as representatives. This is particularly true of warm female members in the known gender treatment. Competence, on the other hand, is relied upon very strongly by groups, even though it is directionally less predictive than warmth of a group representative's success empirically.

V. DISCUSSION

Our paper explores the ways in which gender stereotypes shape group decision-making. We build upon previous work by allowing for free-form chat across group members, providing additional insights into how gender stereotypes operate. We find that women are less likely to be rewarded for their ideas in male-typed domains when gender is known, despite having equal ability and communicating in a similar style. This is partly due to discrimination by fellow group members, and partly due to differences in the propensity to self-promote (particularly when they are in the minority).

The chat data reveal that men and women have very similar styles and contributions to the group on average, as viewed by our blind-to-gender coders. And yet, our coders demonstrate a clear bias in their assessment of member gender, incorrectly believing that warm members are more likely to be female, while more negative members are more likely to be men. Male coders also view more competent members as more likely to be male. Interestingly, while warmth is strongly associated with being a good group representative, groups do not seem to recognize this: warm members, particularly women in the known gender treatment, are less likely to be selected as group representatives. This suggests that stereotypes about communication styles are pervasive, and may shape the expectations for behavior in group decision-making contexts.

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 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. 2018. “A Man’s World? – The Impact of a Male Dominated Environment on Female Leadership,” University of Gothenburg Working Paper in Economics No. 744.

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.

Bordalo, P., Coffman, K. B., Gennaioli N., Schleifer, A. 2018. “Beliefs about Gender,” *American Economic Review*, forthcoming.

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

Catalyst. 2018. “Knowledge Center: Women in S&P 500 Companies,” http://www.catalyst.org.

Chen, J., and Houser, D. 2017. “Gender Composition, Stereotype and the Contribution of Ideas,” GMU Working Paper in Economics No. 17-26.

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.

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

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., Wilmer, J. B. 2012. “Is the Web as Good as the Lab? Comparable Performance from Web and Lab in Cognitive/Perceptual Experiments,” *Psychonomic Bulletin & Review*, 19: 847–857.

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.

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).

Paolacci, G., Chandler, J., Ipeirotis P. G. 2010. “Running Experiments on Amazon Mechanical Turk,” *Judgement and Decision Making,* 5: 411–419.

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.

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).

Table 1: Gender Differences in the Probability of Being Chosen in the Post-Group Stage

Sample	All Chat Treatments		Known Gender Treatment	
	(1)	(2)	(3)	(4)
Female	0.773	1.765	-0.029	-2.747*
	(0.881)	(1.205)	(1.232)	(1.617)
"Maleness" of Question	-0.0002	-0.003	4.408**	4.393*
	(0.011)	(0.012)	(2.221)	(2.259)
Female x "Maleness"	-3.072	2.187	-7.201*	-7.437**
	(2.707)	(3.821)	(3.684)	(3.682)
Gender Known	0.085	1.027		
	(0.237)	(0.948)		
Female x Gender Known		-1.678		
		(1.726)		
Maleness x Gender Known		5.224*		
		(3.040)		
Female x Maleness x Gender Known		-9.142*		
		(5.281)		
Share Female in Group				-3.429
				(2.168)
Female x Share Female in Group				5.440**
				(2.735)
Points in Pre-Group Stage	0.001	-0.0001	-0.003	0.005
	(0.009)	(0.010)	(0.026)	(0.027)
Performance & Demographic Controls	YES	YES	YES	YES
Fixed Effects	YES	YES	YES	YES
Dependent Var. Mean:	33.33	33.33	33.33	33.33
R-squared	0.055	0.058	0.069	0.065
Observations (clusters)	1,656 (276)	1,656 (276)	840 (140)	840 (140)

Notes: Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls include age, student status, race, English language proficiency, income, use of real name, and dummy for whether the US is the country of citizenship and birth; and controls for performance distribution that include difference from maximum group score and difference from average group score. 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 clustered at the group level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table 2: Observed Gender Differences in Chat Behavior

Depedent Variable	Factor 1 ("Competence")	Factor 2 ("Warmth")	Factor 3 ("Negativity")
	(1)	(2)	(3)
Female	0.007	0.0003	-0.0113
	(0.055)	(0.042)	(0.033)
Fixed Effects	YES	YES	YES
Observations (clusters)	1,656 (207)	1,656 (207)	1,656 (207)
R-squared	0.0268	0.0457	0.0685

Notes: Fixed effects include question, round, part, and treatment (gender known or unknown). Robust standard errors clustered at the subject level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table 3: The Effect of Chat Behavior Factors on the Prediction that a Participant is Female

Sample	All	Male Raters	Female Raters
	(1)	(2)	(3)
Factor 1 ("Competence")	-0.016	-0.047***	0.017
	(0.010)	(0.014)	(0.013)
Factor 2 ("Warmth")	0.059***	0.054***	0.069***
	(0.009)	(0.012)	(0.013)
Factor 3 ("Negativity")	-0.051***	-0.0433***	-0.063***
	(0.009)	(0.011)	(0.014)
Rater Was Female	0.084***		
	(0.016)		
Demographic Controls	YES	YES	YES
Dependent Var. Mean:	0.425	0.384	0.472
Observations (clusters)	2,961 (984)	1,584 (526)	1,377 (459)
R-squared	0.039	0.039	0.043

Notes: Rater demographics include gender, education, race, and whether the rater attended high school in the US. Robust standard errors clustered at the rater level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table 4: The Effect of Chat Behavior Factors on Performance in Post-Group Stage

Sample	All Chat Data		Known Gender		Unknown Gender	
	(1)	(2)	(3)	(4)	(5)	(6)
Factor 1 ("Competence")	2.000**	1.918*	2.327*	2.971*	1.852*	1.119
	(0.838)	(1.124)	(1.322)	(1.759)	(1.083)	(1.352)
Factor 2 ("Warmth")	3.200***	3.531**	4.737***	7.904***	2.415	1.519
	(1.114)	(1.722)	(1.448)	(2.074)	(1.697)	(2.664)
Factor 3 ("Negativity")	1.417	2.554	0.764	3.138	2.804	3.470
	(1.435)	(1.949)	(2.043)	(2.798)	(2.010)	(2.733)
Female	0.384	0.364	0.410	0.603	0.566	0.628
	(0.905)	(0.908)	(1.177)	(1.221)	(1.437)	(1.429)
Female x Factor 1		0.211		-0.950		1.271
		(1.624)		(2.504)		(2.028)
Female x Factor 2		-0.557		-5.463*		1.561
		(2.207)		(2.946)		(3.453)
Female x Factor 3		-2.282		-3.687		-1.118
		(2.752)		(3.826)		(3.824)
Gender Known	-0.531	-0.574				
	(1.051)	(1.045)				
Demographic Controls	YES	YES	YES	YES	YES	YES
Fixed Effects	YES	YES	YES	YES	YES	YES
Dependent Var. Mean:	18.01	18.01	18.85	18.85	17.15	17.15
Observations (clusters)	1,656 (207)	1,656	840 (105)	840 (105)	816 (102)	816 (102)
R-squared	0.151	0.152	0.181	0.186	0.186	0.188

Notes: Sample restricted to chat data treatments only. Demographic controls include ranker gender and rankee’s age, student status, race, English language proficiency, income, use of real name, and dummy for whether the US is the country of citizenship and birth. Fixed effects include round, part, and question, as well as treatment (KG) in columns 1 and 2. Robust standard errors clustered at the rater level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table 5: The Effect of Chat Behavior Factors on Ranking by Others in Post-Group Stage

Sample	All Chat Data		Known Gender		Unknown Gender	
	(1)	(2)	(3)	(4)	(5)	(6)
Factor 1 ("Competence")	8.595***	7.655***	8.032***	8.093***	8.872***	6.849***
	(0.811)	(1.074)	(1.224)	(1.594)	(1.107)	(1.512)
Factor 2 ("Warmth")	-1.916**	-1.479	-4.105***	-1.195	-0.468	-1.867
	(0.860)	(1.340)	(1.186)	(2.169)	(1.168)	(1.769)
Factor 3 ("Negativity")	-2.204**	-0.623	-2.990*	-0.268	-2.283	-1.335
	(1.017)	(1.664)	(1.633)	(2.724)	(1.379)	(2.201)
Female	1.189	1.151	0.533	0.664	2.249*	2.316*
	(0.904)	(0.899)	(1.359)	(1.357)	(1.258)	(1.259)
Female x Factor 1		2.000		0.362		3.664*
		(1.380)		(2.005)		(1.885)
Female x Factor 2		-0.907		-5.077*		2.295
		(1.866)		(2.805)		(2.446)
Female x Factor 3		-3.259		-4.787		-1.566
		(2.426)		(3.745)		(3.171)
Gender Known	1.906***	1.831**				
	(0.734)	(0.747)				
Demographic Controls	YES	YES	YES	YES	YES	YES
Fixed Effects	YES	YES	YES	YES	YES	YES
Dependent Var. Mean:	26.03	26.03	26.67	26.67	25.37	25.37
Observations (clusters)	1,656 (207)	1,656 (207)	840 (105)	840 (105)	816 (102)	816 (102)
R-squared	0.114	0.116	0.118	0.122	0.132	0.137

Notes: Sample restricted to chat data treatments only. Demographic controls include ranker gender and rankee’s age, student status, race, English language proficiency, income, use of real name, and dummy for whether the US is the country of citizenship and birth. Fixed effects include round, part, and question. Robust standard errors clustered at the group level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.