GENDER STEREOTYPES IN DELIBERATION AND TEAM DECISIONS[*]

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

Harvard Business School Cornerstone Research Wellesley College

First Draft: November 2018

Current Draft: November 2018

Abstract:

This paper explores the ways in which gender stereotypes shape group decision-making. We design a laboratory experiment that uses a novel task that admits a variety of valid, subjective answers, promoting group discussion. We allow for free-form chat across group members, providing 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. The analysis of chat data by third-party coders reveals pervasive incorrect beliefs that warm participants are more likely to be female, while more critical participants are more likely to be men. Male coders also view more competent participants as more likely to be male. These stereotypes about communication styles may shape expectations for behavior in group decision-making contexts.

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

JEL Classifications: C90, J16, J71

[*] We thank 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

[‡] Cornerstone Research, 699 Boylston St, Boston, MA, USA, CFlikkema@cornerstone.com

[§] 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 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 build on these results 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 in 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 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 warmer participants are less likely to be rewarded for their ideas. 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

IIA. THE TASK

Participants in our experiment play multiple rounds of a *Family Feud* style task.[1] We chose this task because we felt it allowed for ample discussion, reasonable disagreement over what answers might be best, balancing a degree of subjectivity with an objective scoring rule.

In our task, group members view a question, like the one in the example below. The goal is to guess an answer to the question that would be frequently given by others. Importantly, it does *not* matter how many participants *in our experiment* gave a particular answer. The number of points assigned to each answer is equal to the number of Family Feud's own survey participants (out of 100) who gave that particular answer.

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.

We used a set of 8 questions designed to vary in their gender-type (4 female-oriented and 4 male-oriented). 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 a “*maleness*” index which ranges from -0.57 (the most female-typed question) to 0.51 (the most male-typed question), based upon external ratings of the questions from a separate sample of Amazon Mechanical Turk participants.[2]

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, subjects were randomly re-matched in groups of three, using stranger matching.

We vary whether or not gender is revealed to participants. In the unknown-gender treatment, participants were identified in each round by an ID number. In the known-gender treatment, we revealed gender to participants, both by using names and allowing them to hear their fellow group members say “here” in a role call (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*.”

Next, subjects entered the “*group*” stage where they could chat over the computer interface for 60 seconds with each other. 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).

Then, participants 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.” We used one randomly-chosen participant’s ranking to determine the representative: the person ranked first had a 60% chance of being the group representative; second had a 30% chance; third had a 10% chance.

Finally, there was a “post-group” stage where subjects again submitted individual answers to the same 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: 10% chance they were paid for individual answer in pre-group stage, 80% chance paid for group answer given by selected representative, and 10% chance paid for individual answer in post-group stage. In addition, the person selected as the “group representative” received a bonus payment of $2, providing a material incentive to be

[2] These AMT participants rated each of the eight Family Feud questions on a scale of -1 to 1 where -1 was “women know much more” and 1 was “men know much more.” Our maleness scale is the average rating given by the AMT participants. See Appendix F.

chosen. In total 207 subjects participated in our chat treatments, 105 in the known-gender version and 102 in the unknown-gender version in the CLER Lab at HBS.[3]

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 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.

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. 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 D1).

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 to be chosen as the maleness of the question increases, suggestive of

[3] We also conducted two different control treatments. Each participant was randomly assigned to two treatments, one known gender and one unknown gender. Each participant participated in at most one chat treatment. See Appendix A for a summary of treatments and Appendix D for full details on the control treatments and their results.

stereotyping (see Table D2). The combination of differences in self-promotion behavior and discrimination from others leads to the gender gap in serving as group representative.

IV. ANALYSIS OF CONVERSATION DATA

IVA. METHODOLOGY

To make sense of the 276 conversations from our experiment in an objective and tractable way, we recruited 1000 AMT workers to read these conversations and provide impressions of the contributions made.[4]

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 conversation, 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 mostly 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

[4] 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.

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.

[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. 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. 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]

Finally, we ask whether the three factors of competence, warmth, and negativity predict how members were ranked in the experiment *by their fellow group members*. That is, are more competent (warm/negative) participants more (less) likely to be ranked favorably by others? We explore this in Table 4, predicting a participant's rating by another group member from her average rating on each factor for that conversation. First, we pool across the two treatments, presenting the specification without the conversation ratings, juxtaposed next to the specification that includes the coder ratings (Column 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 4 and 7).

Warmth has a negative impact on ranking in the KG treatment (Column 4), but not in the UG treatment (Column 7). Column 5 reveals that only women in the KG treatment are penalized for warmth. On the other hand, men in the KG treatment (Column 5) and men and women in the UG treatment (Column 8) 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.

[TABLE 4 ABOUT HERE]

Note that the inclusion of chat style factors does not fully explain the gender stereotyping in response to the maleness of the question in the main experiment. In other words, the large negative effect for women as maleness of question increases (the coefficient on female x maleness) in the KG treatment in Column 3 is no longer statistically significant in Columns 4 and 5, but remains large in magnitude.

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.

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. 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	0.775	1.744	-0.0286	-2.747*
	(0.877)	(1.195)	(1.232)	(1.617)
"Maleness" of Question	1.561	-1.390	4.408**	4.393*
	(1.412)	(2.133)	(2.221)	(2.259)
Female x "Maleness"	-3.000	2.241	-7.201*	-7.437**
	(2.670)	(3.764)	(3.684)	(3.682)
Gender Known	0.135	1.064		
	(0.235)	(0.960)		
Female x Gender Known		-1.645		
		(1.716)		
Maleness x Gender Known		5.153*		
		(3.017)		
Female x Maleness x Gender Known		-9.164*		
		(5.246)		
Share Female in Group				-3.429
				(2.168)
Female x Share Female in Group				5.440**
				(2.735)
Points in Pre-Group Stage	0.000868	-6.15e-05	-0.00331	0.00451
	(0.00921)	(0.0102)	(0.0257)	(0.0267)
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 for 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")	-0.0159	-0.0470***	0.0172
	(0.00969)	(0.0139)	(0.0133)
Factor 2 ("Warmth")	0.0591***	0.0541***	0.0686***
	(0.00864)	(0.0118)	(0.0131)
Factor 3 ("Negativity")	-0.0509***	-0.0433***	-0.0626***
	(0.00865)	(0.0107)	(0.0140)
Rater Was Female	0.0842***		
	(0.0162)		
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 Ranking by Others in Post-Group Stage

Sample	All Chat Treatments		Known Gender Treatment			Unknown Gender Treatment		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	1.269	1.201	0.771	0.558	0.754	2.156	2.461**	2.526**
	(0.990)	(0.902)	(1.394)	(1.328)	(1.343)	(1.365)	(1.236)	(1.250)
"Maleness" of Question	0.0546	0.612	1.765	0.765	0.936	-3.465	-3.000	-3.120
	(1.863)	(1.767)	(3.030)	(3.026)	(3.050)	(2.783)	(2.600)	(2.672)
Female x "Maleness"	-4.144	-2.971	-8.934**	-6.435	-6.195	1.507	2.089	2.366
	(3.004)	(2.760)	(4.148)	(3.890)	(3.955)	(4.212)	(3.776)	(3.835)
Gender Known	1.193*	1.521**						
	(0.658)	(0.710)						
Factor 1 ("Competence")		7.765***		7.315***	7.477***		7.970***	5.885***
		(0.802)		(1.222)	(1.537)		(1.079)	(1.476)
Factor 2 ("Warmth")		-1.358		-3.616***	-1.106		0.217	-1.026
		(0.883)		(1.176)	(2.107)		(1.257)	(1.824)
Factor 3 ("Negativity")		-1.523		-1.960	-0.235		-2.051	-1.726
		(1.004)		(1.594)	(2.598)		(1.380)	(2.091)
Female x Factor 1					0.0339			3.791**
					(2.034)			(1.806)
Female x Factor 2					-4.452*			1.967
					(2.757)			(2.410)
Female x Factor 3					-2.659			-0.541
					(3.658)			(2.997)
Points in Pre-Group Stage	-0.0770**	-0.151***	-0.0993**	-0.142***	-0.148***	-0.0444	-0.138**	-0.136**
	(0.0353)	(0.0370)	(0.0484)	(0.0536)	(0.0540)	(0.0620)	(0.0629)	(0.0627)
Performance & Demographic Controls	YES	YES	YES	YES	YES	YES	YES	YES
Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES
Dependent Var. Mean:	26.03	26.03	26.667	26.667	26.667	25.368	25.368	25.368
Observations (clusters)	1,656 (276)	1,656 (276)	840 (140)	840 (140)	840 (140)	816 (136)	816 (136)	816 (136)
R-squared	0.055	0.123	0.069	0.127	0.130	0.070	0.146	0.151

Notes: Sample restricted to chat data treatments only. Controls for performance include individual points, difference from maximum group score, and difference from average group score in pre-group stage. Note that performance information was not observable in the Chat treatment. 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 and part. Robust standard errors clustered at the group level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.