Who Does the Talking Here? The Impact of Gender Composition on Team Interactions*

David Hardt†, Lea Mayer‡, Johannes Rincke§

January 10, 2024

Abstract

We analyze how the gender composition of teams affects team interactions. In an online experiment, we randomly assign individuals to gender-homogenous or gender-mixed teams of four. Teams meet in an audio chat room and jointly work on a team task that is gender-neutral in terms of individual productivity. By design, effects on team performance can only work through communication. We find that all-male teams communicate more than all-female teams and outperform teams of both alternative gender compositions. In mixed teams, men strongly dominate the team conversation quantitatively. Considering the ranking in terms of communication shares in mixed teams, we document that high-skilled men talk the most, followed by low-skilled men. High-skilled women talk more than low-skilled women, but much less than low-skilled men. We conclude that the gender composition significantly impacts the way how teams exchange information.

Keywords: Teams; teamwork; gender composition; communication; team performance; online experiment

JEL Codes: C92, C93, D83, J16

*The authors would like to thank Alexander Cappelen, Florian Englmaier, Ulrich Glogowsky, Emanuel Hansen, Lea Heursen, Dorothea Kübler, John List, Markus Nagler, Hessel Oosterbeek, Dan-Olof Rooth, Alicia von Schenk, David Schindler, Amelie Schiprowski, Hanna Schwank, Boris van Leeuwen, Fabian Waldinger and three anonymous reviewers for helpful comments. Seminar participants at various places provided valuable feedback. Rincke gratefully acknowledges financial support from the German Research Foundation (RI 1959/5-1) and the Hans-Frisch Foundation. This experiment was pre-registered at the AEA Social Science Registry as AEARCTR-0007989, `https://www.socialscienceregistry.org/trials/7989`.

†University of Erlangen-Nuremberg; david.hardt@fau.de

‡University of Erlangen-Nuremberg; lea.mayer@fau.de

§Corresponding author; University of Erlangen-Nuremberg and CESifo; johannes.rincke@fau.de

1 Introduction

Two powerful trends have recently transformed how companies and other organizations operate: The rise of teamwork and a tendency towards increased gender diversity in traditionally male-dominated domains. Teamwork has become pervasive in the workplace, and the labor market increasingly rewards workers for their collaboration, communication, and leadership skills (Deming, 2017; Weidmann and Deming, 2021; Edin et al., 2022).[1] At the same time, many workers tend to collaborate in increasingly gender-diverse environments. For instance, the share of women among STEM workers steadily increased over the past decades, reaching 50 percent in 2019 (Pew Research Center, 2021). Similarly, the most recent data show that the share of women on Fortune 500 and S&P 500 company boards almost doubled over the past 10 years, reaching 26.5 and 30.6 percent, respectively (Spencer Stuart, 2021; Alliance for Board Diversity and Deloitte, 2021).[2]

The overlapping of both trends has led to a significant rise in gender-mixed teamwork. Some dimensions of this structural change in how organizations operate have been extensively analyzed, including the benefits and costs of board diversity (for a review, see Adams et al., 2015). Other aspects of the shift towards gender-mixed teamwork have received much less attention, including the question of how gender diversity impacts social interaction in teams, in particular communication. Evidence on how a change in the team gender composition affects the behaviors of individual team members in these dimensions is particularly scarce.

This paper presents experimental evidence on how a team's gender composition affects team interactions. In the experiment, randomly composed teams of four met in an online audio chat room to jointly work on a non-routine team task. The fact that team members had to collaborate differentiates our approach from designs where team members work on their tasks separately, and team members' performance in these different tasks then determines the joint output (Ivanova-Stenzel and Kübler, 2011). The team task consisted of a series of single-choice problems on business cases, and the design made sure that solving problems required communication among team members. Our main outcomes include measures of communication at individual

[1]Recent figures suggest that almost 80 percent of U.S. employment is in occupations where teamwork is judged either a "very" or "extremely" important part of the job (O*NET OnLine, 2022), and employers tend to consider teamwork as one of the most important attributes of new employees. Earlier sources discussing the rise of teamwork include Lazear and Shaw (2007) and Owan (2014).

[2]Increased gender diversity also affects domains outside firms. In the U.S. House of Representatives, the share of seats held by women doubled in the last 20 years, reaching 28.3 percent in 2020 (Congressional Research Service, 2022). The share of women in U.S. Cabinet-level positions reached 48 percent under President Biden, more than four times higher than 40 years ago (Center for American Women and Politics, 2022).

and team level and team performance. In addition, we explore how past exposure to gender-mixed teamwork affects preferences and beliefs about further teamwork.

Based on a sample of 1368 subjects in 342 teams, we derive four sets of main results. First, all-male teams communicate more than mixed and all-female teams. These differences are more pronounced if we consider only words that are topically related to the problems the teams discuss ("topic words"). Second, all-male teams outperform both gender-mixed and all-female teams. An exploratory analysis suggests that team performance is driven by the usage of topic words. Third, men and women behave very differently in gender-mixed relative to gender-homogenous teamwork. Whereas men in mixed teams speak significantly more than men in all-male teams, women adjust their communication behavior in the opposite direction. As a result, mixed-team communication is characterized by a sizeable gender gap, with men uttering almost 70 percent more words than women. Finally, analyzing gender vs. cognitive skills as determinants of communication shares in mixed teams, we find that high-skilled men talk the most, followed by low-skilled men, whose communication shares are only slightly lower and not statistically different from those of high-skilled men. High-skilled women talk significantly less than low-skilled men and significantly more than low-skilled women.

We recruited our subjects via an online platform at a large public university in Germany. The platform allows us to access rich individual background data, including students' gender and A-level GPA as a comprehensive measure of cognitive skills. Exploiting this feature, we recruited samples of female and male subjects that were balanced in cognitive skills.

Our main contribution is clean experimental evidence on gender differences in team communication. Since the sharing and processing of information is fundamental for translating team-level skill diversity into productivity gains (e.g., Hamilton et al., 2012; Deming, 2017; Lyons, 2017; Weidmann and Deming, 2021), we provide new evidence on a likely channel through which the team gender composition may ultimately affect team performance in many real-world settings. In fact, our data suggest that the ability of all-male teams to consistently outperform teams of alternative composition is driven by all-male teams communicating more than mixed and all-female teams.[3]

In terms of generalizability, one would optimally want to implement a natural field experiment and thereby avoid issues like selection into participation and possible scrutiny effects (Al-Ubaydli and List, 2013). However, we firmly believe that for our research question, an online experiment is very useful. Most importantly, in our sample we can credibly rule out gender differences in task-specific ability, something

[3]Since the teams were composed of strangers, our findings cannot be explained by differences in group cohesion (Gächter et al., 2023).

that would be difficult to achieve in a natural setting. In terms of selection, our subjects are broadly representative of the overall student population at the university at which we implemented our study. Regarding attrition, we document that attrited and non-attrited teams have similar characteristics. Considering the naturalness of the task, setting, and time frame, we put subjects in a situation that has some similarity to a real-world team task where subjects collaborate for a short period of time with strangers, using verbal communication to coordinate and exchange ideas. Further research is needed to study if our insights transfer to settings with repeated interaction, tasks that are less artificial than our business cases, and less scrutiny. Regarding scalability, we would like to stress that our subjects are used to collaborating in more or less gender-diverse settings from a myriad of group assignments. The fact that we observe very strong treatment effects on communication among these subjects suggests at least some scalability of our findings.

The paper relates to several literature strands. As regards team performance, several papers study the effect of women on corporate boards. Whereas some papers report positive effects on firm performance (Campbell and Mínguez-Vera, 2008; Terjesen et al., 2016), others find none (Chapple and Humphrey, 2014) or even negative effects (Adams and Ferreira, 2009; Ahern and Dittmar, 2012; Matsa and Miller, 2013). Among the experimental studies, Hoogendoorn et al. (2013) consider ventures started by undergraduate business students and show that sales and profits increase when the share of women increases from a low to an intermediate level. Contrasting evidence on the effect of team gender composition includes Lamiraud and Vranceanu (2018), who demonstrate that all-male and mixed teams in a student business game perform significantly better than all-female teams. Apesteguia et al. (2012) obtain similar results. Marx et al. (2021) find that gender-homogeneous teams perform significantly better than gender-diverse teams.[4] We provide the first study that focuses on communication as a likely channel through which the team gender composition may affect team performance.

The paper also relates to literature on gender differences when people speak in public. Regarding style and tone, several studies document that men often establish dominance over women through hostility and interruptions (Jacobi and Schweers, 2017; Dupas et al., 2021; Miller and Sutherland, 2023). Interestingly, in our experiment, we find no support for these channels. As regards the quantity of communication, observational data suggest that in public settings like academic conferences and seminars, women tend to ask fewer questions than men (Davenport et al., 2014; Hinsley et al., 2017; Carter et al., 2018; Dupas et al., 2021). This might have to do

[4]For further references, see Azmat and Petrongolo (2014). Further dimensions of team diversity are discussed in, e.g., Hoogendoorn and Van Praag (2012), Hamilton et al. (2012), Hjort (2014), Lyons (2017), and Marx et al. (2021).

with women having a stronger aversion to speaking in public, but the experimental evidence on this question is mixed (De Paola et al., 2021; Buser and Yuan, 2023). Regarding communication behavior in small groups, studies in psychology tend to find that men dominate women in terms of speaking time (MacLaren et al., 2020). Considering classroom interaction, boys tend to participate more, initiate contact more often with the teacher, and interrupt more than girls (Kelly, 1988). These gender gaps seem to be socially acquired (Aukrust, 2008), a conclusion that is in line with our findings. We advance this literature by the first systematic analysis of style and quantity of communication in teams that vary exogenously in their gender composition.[5] Furthermore, we can rule out gender differences in individual determinants of communication behavior, such as ability and experience.

The literature has also studied individuals' aspirations to lead. Consistently, men are found more willing to lead than women (Ertac and Gurdal, 2012; Arbak and Villeval, 2013; Born et al., 2022; Haegele, 2022), and this gender gap seems to be socially acquired (Alan et al., 2020). Individuals use speaking time to express leadership, and infer emerging leadership from how much other individuals talk (Schmid Mast, 2002; MacLaren et al., 2020). Compared to women on majority-women teams, token women are perceived as less influential by peers and are less likely to be chosen as group leaders (Karpowitz et al., 2023). We add to this literature by showing that in gender-mixed teams, men are much more likely than women to quantitatively dominate the team conversation, suggesting stronger leadership aspirations. Ultimately, gender difference in the tendency to dominate a team conversation may be part of the explanation of why women are still strongly under-represented in real-world leadership positions (e.g., Bertrand and Hallock, 2001; Blau and Kahn, 2017).

Regarding preferences for teamwork, the literature has mainly discussed worker heterogeneity in the decision to join teams (Hamilton et al., 2003; Bandiera et al., 2013; Cooper et al., 2021). Among the studies addressing gender, Kuhn and Villeval (2015) use a design where subjects can choose between individual incentives and revenue sharing. They find that women's more optimistic assessments of their prospective teammate's abilities make them more likely than men to choose team-based pay. In contrast to Kuhn and Villeval (2015), we find no difference in how women and men assess the other subject's ability and no overall gender difference in preferences for teamwork. Dahl et al. (2021) show that in a traditionally male-dominated context, men's attitudes towards gender-mixed teamwork are malleable at least in the short

[5] Woolley et al. (2010) study randomly composed teams, but treat communication as an independent variable in an effort to explain group intelligence. Charness et al. (2020) show congestion effects in team communication, but the design does not aim at identifying the effect of team gender composition.

run. Studying a naturally gender-diverse context, we find suggestive evidence that exposure to mixed teamwork affects attitudes.

We pre-registered the experimental design and the data analysis.[6] We explicitly mention in the paper any deviation from the pre-specified analysis. The remainder of the paper is organized as follows. Section 2 explains the setting and the experimental design. Section 3 elaborates on the data and the empirical strategy. Section 4 discusses the results, and Section 5 concludes.

2 Setting and Experimental Design

2.1 Online Platform and Subject Pool

We implemented our framed field experiment using an online platform at the University of Erlangen-Nuremberg, Germany, with about 10,000 registered users.[7] It works similarly to other online panels in which individuals can register to work on and get paid for short tasks. Our key advantage is that we can link the experimental data to the university's registry data. This data contains age, gender, field of study, and A-level GPA. The GPA is the grade of the students' university entrance certificate earned at high school. We use the A-level GPA as a proxy for cognitive skills.[8][9] To each of the 23 sessions, we invited (for a fixed date and time) a random subsample of subjects from the pool via email, stratified by gender and cognitive skills. The email informed subjects that a quiet working space with a stable internet connection and a device with a microphone were prerequisites for participation.

2.2 Experimental Design

Overview The experiment had two stages. In stage 1, subjects worked on a real-effort team task. Randomly composed teams of four subjects met in an online audio chat room to jointly work on a series of 10 single-choice problems related to two business cases.[10] Stage 2 consisted of a choice experiment. We elicited preferences over future teamwork and various beliefs, conditional on random variation in two dimensions:

[6] `https://www.socialscienceregistry.org/trials/7989`

[7] We programmed the experiment with oTree (Chen et al., 2016).

[8] The fact that the A-level GPA combines grades obtained from a large variety of assessments (instead of grades from a specific form of assessment) should alleviate concerns regarding gender gap variation across assessment types (Borghans et al., 2016; Graetz and Karimi, 2022).

[9] We invited only subjects younger than 32 years and fluent in German. Data collection took place in 2021. We ran pilot sessions between March and July. In these sessions, we tested the functionality of our webpage and the invitation procedure. The experimental sessions were conducted between the end of July and November.

[10] See Online Appendix Section F for further details.

the gender composition of a subject's team in stage 1, and the gender of a subject's teammate in possible further teamwork in stage 2. Figure B.1 presents a timeline.

Online Environment and Initial Instructions In order to participate, subjects had to log in to their platform account at the time communicated in the invitation email. The webpage informed subjects that they were about to participate in a research project on human interaction in groups that would involve an audio chat with other participants. The webpage asked for consent to record the audio chat for research purposes and to link background information on the subject to the experimental data. After a microphone test, the webpage redirected the subjects to a screen with instructions. The instructions informed subjects that they would earn a fixed show-up fee of €10, that the session would consist of two parts, that in the first part, they would work with three other randomly selected participants on a team task, and that an audio chat room would enable communication between team members. The instructions also explained that the team task in the first part would consist of 10 problems and that each team member would earn a bonus of €1 per problem conditional on all team members marking the correct answer on their screen individually. The bonus scheme ensured that all team members had incentives to coordinate on joint answers. We here rely on Englmaier et al. (2023), who show that incentives improve team performance in non-routine analytical team tasks.

Stage 1 of the Experiment At the beginning of stage 1, the subjects were randomly assigned to teams of four. The teams' gender composition varied between all-male, mixed (two women and two men), and all-female. For randomization, we used registry data (e.g., gender and A-level GPA). The scheme ensured that each team consisted of two subjects with above-median and two subjects with below-median A-level GPA (for details on the sampling frame, see Section 2.3). Subjects who could not be assigned to a team received a show-up fee and were re-invited to later sessions.

The webpage then redirected the subjects to a team-specific browser-based audio chat room (no video). In the chat room, the subjects were (randomly) labeled from 1 to 4. Each team member's number was shown as an avatar, and the avatar currently speaking was highlighted (for screenshots, see Section E of the Online Appendix). This enabled subjects to infer who was speaking and address each other. The teams were given time to familiarize themselves with the chat room and discuss the team-task instructions together. The instructions explained that the teams had three minutes to work on each problem and to coordinate on a solution.[11]

[11]The instructions explained that the bonus for a given problem would only be paid conditional on all team members marking the correct answer on their own screen before the three-minute countdown

From here on, the webpage directed the subjects through the team task. The task was divided into two blocks of five problems each. Each block featured a business case that we adapted from publicly available training materials provided by the HR department of a large international strategy consultancy. Each case involved extensive information material (text plus a table or chart). The subjects were given extra time to study the material whenever new material was shown (i.e., the reading time did not count towards the three minutes available for each problem). Whenever a new problem started, the webpage displayed four written statements, one of which was true. Subjects then had three minutes to discuss the problem and mark one statement as the team's solution to the given problem. The timing of the experiment was fixed, and all subjects in a team were redirected to a given page at the same time.

Stage 1 of the experiment ended with a farewell screen. The subjects then filled out a survey individually.[12] The survey elicited perceptions about the team task and the team's communication. The subjects also stated their perception of how many of the other team members were female.[13]

Stage 2 of the Experiment At the beginning of stage 2, we topped up the fixed payoff by €2 for completing the experiment. The subjects then met another randomly selected subject in the audio chat room for one minute. The matching algorithm made sure that all subjects met a stranger (i.e., a subject from a different first-stage team). The purpose of letting pairs of subjects meet in the audio chat room was to enable the subjects to learn about the gender of a randomly drawn other subject in a way that would not reveal our interest in gender-related preferences or beliefs (for details on the matching procedure, see Section 2.3). In the chat room, each subject saw on the screen a private numerical five-digit key, together with an empty input field. The subjects' task was to exchange their keys and enter the other subject's key into the input field. The request to exchange the private keys made sure that the subjects talked to each other, thereby enabling both subjects in a pair to make inferences about the other subject's gender. When the time allocated to the key-exchange task was over, the audio chat closed and subjects could no longer communicate with each other.[14] Subjects who could not be assigned to a pair were informed that no matching partner was available for them. For these subjects, the following elicitation of preferences and beliefs was skipped, and the subjects were redirected to the final survey page.

---

for the respective problem expired, and that the session would be closed for the whole team if someone would leave the session for more than 90 seconds.

[12]When working on the survey, the audio chat room was closed.

[13]We embedded this item in obfuscation questions.

[14]To be able to elicit the Big 5 personality traits in the final survey from all subjects, we let subjects proceed even if they did not enter the correct key. However, we exclude these subjects from the estimation sample of stage 2 (for details on estimation samples, see Section 2.3).

Once the audio chat had closed, we informed the subjects about the possibility that they would work on another task similar to the one in stage 1 for 15 minutes, and asked subjects to state their preference for working on the task individually or in a two-person team with the subject they had met in the audio chat room. To elicit preferences over teamwork, we used the following mechanism: Before the subjects stated their preference, we informed them about a random draw with three possible outcomes: (a) both subjects would not work on the task at all; (b) both subjects would work on the task individually irrespective of their stated preference; and (c) their stated preferences would be implemented as follows: they would work as a team if they both indicated this as their preferred option, and they would both work individually otherwise. Subjects knew that, in case of individual work, they would earn a bonus of €1 for each correct answer, and that in case of teamwork, they would earn the same bonus for each problem answered correctly by both teammates (same procedure as in first-stage team task). We did not communicate a probability distribution over the different possible outcomes. The implemented probabilities were 90 percent for outcome (a) and five percent for outcomes (b) and (c), respectively. As a result, the mechanism to elicit preferences over teamwork was incentive-compatible, but it also made sure that the majority of subjects did not have to do another task.[15] We also elicited the subjects' beliefs about their own and the other subject's productivity when working on the task individually, and team productivity when working with the other subject. To elicit these beliefs, we asked the subjects to imagine a task similar to the one in the first stage, but comprising 20 problems. In addition, we elicited beliefs about team communication and team interaction in the hypothetical case that both subjects would work as a team. Stage 2 of the experiment ended with survey questions. The survey asked the subjects about their perception of whether the person they met in the chat room was female and elicited the Big 5 personality traits following Gerlitz and Schupp (2005).[16] We then implemented the random draw regarding the task, and (if determined by the draw) subjects worked on the task (individually or as a team).

2.3 Sampling, Attrition, and Balancing Checks

Formation of Teams in Stage 1 Our sampling procedure aimed at symmetry in the team-level composition of cognitive abilities across first-stage teams of different gender compositions. For that purpose, when randomly assigning subjects to teams, each

[15] To address concerns that the outcome would reveal that one had rejected the other subject as a potential teammate (or had been rejected by the other subject), we pointed the subjects to the fact that even if both subjects would end up working alone, this would not reveal their stated preferences.

[16] The question on the other subject's gender was again embedded in obfuscation questions.

team was formed by drawing two subjects of above-median and two of below-median skills, measured by A-level GPA.

Random Assignment of Subjects to Potential Teammates in Stage 2 In stage 2, subjects were randomly assigned to a potential teammate from another first-stage team. First, we randomly formed pairs of first-stage teams. Second, we randomly matched the subjects from a given pair across teams into pairs of subjects.[17]

Sample Size, Attrition, and Balancing Checks In the pre-analysis plan, we committed to collect data on between 200 and 400 first-stage teams.[18] We stopped the data collection when we had exhausted the subject pool by repeatedly inviting subjects who had not responded before. In total, 411 teams took part in the experiment. 69 teams attrited during their session, leaving us with a sample of 342 teams who finished the team task (114 all-male, 113 mixed, and 115 all-female teams). Attrition was mainly due to teams being disqualified when individual team members dropped out during the team task (54 teams). We treat another 15 teams as attrited where individual members seemed to experience unforeseen technical issues, like problems unmuting their microphone.[19] Table 1 reports balancing checks at the team level for non-attrited teams and shows that all-male, mixed, and all-female teams were balanced in observable team characteristics. Online Appendix Table A.1 documents that attrited and non-attrited teams have similar characteristics.

Table A.2 in the Online Appendix documents attrition in stage 1 at individual level. With 342 non-attrited teams, our estimation sample at the individual level comprises $342 \times 4 = 1368$ observations. Table 2 reports balancing checks at the individual level for stage 1 by comparing women and men between gender-homogenous and mixed teams. Apart from male engineering students being over-represented in all-male relative to

[17] If the number of first-stage teams in a session was odd, we randomly selected three first-stage teams, then randomly selected six subjects from those teams, and randomly assigned each of them one of the remaining subjects from another team. With all remaining first-stage teams, we proceed as described before. Figure B.2 in the Online Appendix illustrates the matching.

[18] During the pilot sessions, we tried to increase the efficiency of data collection by adjusting the invitation procedure in a stepwise manner (text of invitation email, timing of sessions and reminder emails, number of subjects invited, etc.). We managed to reach a participation rate of about 10 percent, but when we pre-registered the design, we did not know how participation rates would evolve over time (i.e., when repeatedly inviting subjects who had not responded to an invitation before).

[19] When transcribing the teams' audio files, we became aware that 46 teams had individual members who did not contribute at all to the team conversation. Our design did not prevent subjects from staying silent throughout the team task, and we have no means to objectively determine whether silent subjects experienced unforeseen technical issues, like problems unmuting their microphone, or actively decided not to contribute to the team conversation. To account for this issue, we drop teams with silent members in which team members gave identical answers in less than five problems. The rationale for this rule is that, if a team managed to coordinate, it is likely that silent members could at least hear the team conversation. Teams in this situation would effectively work as teams with three active and one passive member and would still be able to earn a bonus.

Table 1: Balancing Checks, Team Level

	All-male teams (1)	Mixed teams (2)	All-female teams (3)	p-value all equal (4)
Mean A-level GPA	2.73	2.74	2.76	0.30
	(0.17)	(0.16)	(0.17)	
Maximum A-level GPA	3.45	3.47	3.43	0.59
	(0.30)	(0.26)	(0.31)	
Minimum A-level GPA	2.00	2.03	2.06	0.37
	(0.31)	(0.30)	(0.30)	
Share top-tier high school	0.83	0.81	0.84	0.51
	(0.19)	(0.19)	(0.19)	
Mean age	22.71	22.79	22.56	0.49
	(1.60)	(1.41)	(1.49)	
Maximum age	26.32	26.50	25.76	0.08
	(2.90)	(2.55)	(2.44)	
Minimum age	19.71	19.67	19.77	0.88
	(1.56)	(1.54)	(1.47)	
Share foreign nationality	0.04	0.03	0.04	0.42
	(0.10)	(0.08)	(0.09)	
N. of obs.	114	113	115	342

Notes: This table reports team-level balancing checks. Columns (1) to (3): Means and standard deviations. Column (4): p-values for tests of hypothesis that all three means are equal.

mixed teams, individual characteristics of women and men are balanced between teams of different gender compositions. In terms of selection into participation, we also compared our subjects to the overall student population. We found our sample to be representative in terms of GPA, age, gender, type of university entrance certificate, and nationality (results available upon request).

Of the subjects who did not attrit in the first stage, 960 subjects entered the second stage of the experiment.[20] 229 subjects attrited during stage 2, leaving us with a sample of 731 subjects. Attrition during stage 2 happened when subjects could not be matched to another subject,[21] did not enter the correct keys when meeting in the audio chat room or skipped preference and/or beliefs elicitation questions in stage 2. We did not exclude subjects in stage 2 from the experiment and elicited the Big 5 personality traits in the final survey from all subjects present at that stage. Table A.3 in the Online Appendix documents attrition in stage 2. Tables A.4 and A.5 in the Online Appendix report balancing checks for the sample of subjects who finished stage 2.

[20]The fact that the number of subjects entering stage 2 is lower than the number of subjects finishing stage 1 is due to subjects dropping out between the stages and due to the pilot sessions being part of our data. These sessions did not include stage 2.

[21]Subjects who dropped out between the stages were missing in the second-stage matching. Hence, if one subject dropped out between the stages, this left another subject without a matching partner.

Table 2: Balancing Checks, Individual Level

	Males assigned to			Females assigned to		
	All-male teams (1)	Mixed teams (2)	p-value both equal (3)	All-female teams (4)	Mixed teams (5)	p-value both equal (6)
A-level GPA	2.73	2.75	0.71	2.76	2.73	0.52
	(0.63)	(0.62)		(0.60)	(0.62)	
Top-tier high school	0.83	0.82	0.79	0.84	0.81	0.27
	(0.38)	(0.39)		(0.36)	(0.39)	
Age	22.71	22.62	0.74	22.56	22.97	0.10
	(3.28)	(3.20)		(2.94)	(3.20)	
Foreign nationality	0.04	0.03	0.58	0.04	0.02	0.20
	(0.19)	(0.16)		(0.20)	(0.16)	
Study program: Master level	0.28	0.24	0.27	0.21	0.24	0.44
	(0.45)	(0.43)		(0.41)	(0.43)	
Study program: Arts and humanities	0.19	0.21	0.51	0.29	0.27	0.69
	(0.39)	(0.43)		(0.45)	(0.43)	
Study program: Engineering	0.28	0.19	0.01	0.13	0.14	0.81
	(0.45)	(0.37)		(0.34)	(0.37)	
Study program: Natural sciences	0.10	0.12	0.46	0.10	0.10	0.80
	(0.30)	(0.31)		(0.29)	(0.31)	
Study program: Economics and business	0.30	0.32	0.55	0.28	0.26	0.55
	(0.46)	(0.45)		(0.45)	(0.45)	
N. of obs.	456	226	682	460	226	686

Notes: This table reports subject-level balancing checks. Columns (1), (2), (4), (5): Means and standard deviations. Columns (3) and (6): p-values for tests of hypothesis that the means are equal.

3 Empirical Strategy

In this section, we describe how we estimate the effects of interest and explain our primary outcomes.

3.1 Team Level Estimations

We derive our team-level results from the estimation equation

$$Y_g = \beta_0 + \beta_1 T1_{FM,g} + \beta_2 T1_{FF,g} + X_g'\gamma + u_g \tag{1}$$

where Y_g captures the respective outcome for team g, $T1_{FM,g}$ is an indicator for gender-mixed teams, and $T1_{FF,g}$ is an indicator for all-female teams (all-male teams are the omitted category). X_g captures a vector of team controls. The inclusion of controls is motivated by the fact that gender is a fixed individual attribute that correlates with other individual characteristics. As a result, the random assignment of subjects to teams does not ensure that the team gender composition is orthogonal to team-level means of these characteristics. Following our pre-analysis plan, we account for this fact by including team averages of A-level GPA[22] and age, maximum and minimum A-level GPA and age, the share of team members who graduated from the top-tier high school type ("Gymnasium"), the share of team members with

[22]A-level GPA is coded from 1 (pass) to 4 (best possible grade).

foreign nationality, the share of team members studying at Master level, and a series of variables capturing the shares of team members studying in one of the main study fields (arts and humanities, engineering, natural sciences, and economics/business administration).[23] In addition to the pre-specified covariates, we include an indicator for the presence of a silent team member (see Footnote 19 for details).

We estimate the coefficients in equation (1) by OLS and report robust standard errors. In addition to standard inference, we account for multiple hypothesis testing by reporting p-values that correct for family-wise error rates. We follow the methodology of Barsbai et al. (2020), a generalization of List et al. (2019).

3.2 Individual Level Estimation, Stage 1

To obtain the individual-level results for the first stage of the experiment, we estimate

$$Y_i = \beta_0 + \beta_1 F_i + \beta_2 T1_{FM,i} + \beta_3 F_i \times T1_{FM,i} + X_i' \gamma + u_i, \tag{2}$$

where Y_i denotes the respective outcome for subject i, F_i is an indicator for female subjects, $T1_{FM,i}$ is an indicator for subjects assigned to a gender-mixed team, and X_i captures individual-level controls. We estimate the coefficients by OLS and report robust standard errors accounting for team-level clusters.[24] We include as controls A-level GPA, age, an indicator for subjects who graduated from the top-tier high school type, an indicator for foreign nationality, an indicator for Master students, and indicators for the main fields of study. In addition to the pre-specified covariates, we include an indicator for subjects who worked in teams with a silent member (see Footnote 19 for details). Equation (2) allows us to investigate how (conditional on covariates) exposure to gender-mixed teamwork in stage 1 interacts with a subject's gender in determining stage 1 outcomes.

To analyze patterns in individual communication over time (i.e., across the 10 problems of the team task), we use panel estimations that allow us to derive problem-specific estimates of the interaction effect between F_i and $T1_{FM,i}$. These regressions use subject-by-problem panel data and are based on the equation

$$\begin{aligned} Y_{i,p} &= \alpha + \sum_{p=2}^{10} \beta_p P_p + \sum_{p=1}^{10} \delta_p F_i \times P_p \\ &+ \sum_{p=1}^{10} \eta_p T1_{FM,i} \times P_p + \sum_{p=1}^{10} \theta_p F_i \times T1_{FM,i} \times P_p + X_i' \gamma + u_{i,p}, \end{aligned} \tag{3}$$

[23] The omitted category for field of study is Law/Medicine.

[24] As in the team-level regressions, whenever appropriate we account for multiple hypothesis testing by reporting p-values that correct for family-wise error rates (Barsbai et al., 2020).

where $Y_{i,p}$ captures the outcome of interest of subject i in problem $p = 1, \ldots, 10$, and P_p is an indicator for problem p.

3.3 Individual Level Estimation, Stage 2

Following the pre-analysis plan, the analysis of the individual data from the second stage of the experimental design focuses on identifying the effect of exposure to a gender-mixed team in stage 1. For each primary outcome, we estimate three different equations. First, we estimate equation (2) without the interaction effect. The respective OLS regressions allow us to analyze whether (conditional on covariates) second-stage outcomes differ between subjects who were exposed to gender-mixed teamwork in stage 1 and subjects who were not. Second, we estimate equation (2) including the interaction effect. The respective OLS regressions allow us to analyze how exposure to gender-mixed teamwork in stage 1 interacts with a subject's gender in determining stage 2 outcomes. Third, we estimate (separately for women and men) the equation

$$Y_i = \beta_0 + \beta_1 T2_{F,i} + \beta_2 T1_{FM,i} + \beta_3 T2_{F,i} \times T1_{FM,i} + X_i' \gamma + u_i, \tag{4}$$

where Y_i denotes the respective second-stage outcome for subject i, $T2_{F,i}$ is an indicator for subjects assigned to a female potential partner in stage 2, and $T1_{FM,i}$ is (as before) an indicator for subjects who were assigned to a mixed team in stage 1. The vector of controls X_i is identical to the individual-level estimations for stage 1. Estimating equation (4) separately for women and men informs us about how past exposure to gender-mixed teamwork in stage 1 interacts with the prospective teammate's gender in determining stage 2 outcomes.

To account for possible correlation in second-stage residuals resulting from the interaction among potential second-stage teammates in the audio chat room, all estimations using second-stage outcomes account for clusters that comprise all subjects from the respective first-stage teams. For instance, if the crosswise matching comprised the subjects from first-stage teams j and k, all subjects from teams j and k form one cluster (see Figure B.2 in the Online Appendix for an illustration). We follow Barsbai et al. (2020) to account for multiple hypothesis testing.

3.4 Descriptives

Tables A.6 and A.7 in the Online Appendix display descriptives on outcomes at individual and team level, respectively. The quantitative measures are based on transcripts of recordings capturing the teams' communication. We use the number of words and the number of turns, respectively.[25] On average, subjects contribute 487 words (36.9 turns) to the team conversation in stage 1. The average weight of positive (negative) vocal sentiment is 0.39 (0.26). The perceived positivity and cooperativeness of team communication and the likeability of the task are quite high on average, with the means ranging from 4.0 (likeability) to 4.7 (cooperativeness) on the 5-point Likert scale. In stage 2, 80 percent of subjects prefer teamwork over individual work. On average, the subjects believe they would solve 11 out of 20 problems when working alone. The potential partner is believed to be slightly more productive on average (12.1 problems). Subjects also believe that team productivity would be higher (14.7 problems). The average beliefs regarding positivity and cooperativeness of team communication and the likeability of the task in further teamwork with the potential partner take values between 4.1 (likeability) and 4.5 (cooperativeness). Table A.7 shows that the teams solve 4.4 problems on average. Figure B.3 displays a histogram of team performance.

Because a substantial part of our analysis is concerned with effects on the quantity of communication in stage 1, for illustration purposes, Figure B.4 in the Online Appendix plots the association between the number of words and total speaking time. As expected, the association is very close, with some outliers driven by noise in the audio files.[26] On average, subjects talked for about 3:20 minutes during the team task. The conversation of the average team thus lasted for about 13:00 minutes, with considerable heterogeneity across teams. The fact that the average team used only a fraction of the 30 minutes available for communication likely reflects the type of the team task, which required subjects to process (and potentially re-read) extensive information material.

3.5 Design Checks

Awareness of Team Gender Composition, Stage 1 Table A.8 in the Online Appendix reports a regression of equation (2) using as dependent variable an indicator for subjects whose response to the respective survey item indicates that they were aware of their team's exact gender composition, measured by the number of female team members. Overall, 94 percent of subjects answered the question correctly. The rate of incorrect responses is higher among women in mixed teams. A closer inspection of the data reveals that this is likely due to the framing of the question making it somewhat

[25]We define a turn to be a conversational contribution consisting of at least three words. Adding turns consisting of one or two words to the turn count leads to similar results, but some of our estimates become less precise.

[26]We measure speaking time based on an algorithm that removes periods of silence within turns from recordings and aggregates the remaining time at speaker and team level, respectively. Speaking time tends to be overstated in case of background noise.

more difficult for women to answer correctly.[27] This conjecture is corroborated by the absence of a gender gap if we adjust the awareness indicator to account for this difference (Table A.8, Column 2). 96 percent of the subjects were aware of whether their team was composed only of subjects of the same sex or gender-mixed.

Awareness of Potential Partner's Gender Composition, Stage 2 Table A.9 in the Online Appendix shows that after meeting the potential partner in stage 2, 98 percent of all subjects were aware of their partner's gender, with negligible differences among women and men.

No Gender Differences in Individual Productivity A crucial assumption of our analysis is that the team task itself did neither favor women nor men in terms of individual productivity. In the following, we will discuss several pieces of evidence suggesting that this assumption holds. Part of the evidence is based on a sample of 296 subjects who worked on the exact same task as the teams but under an individual piece rate.[28] Online Appendix Table A.10 provides balancing checks. Online Appendix Table A.11 reports an OLS regression of the number of problems solved on individual characteristics. The gender difference in performance is small and insignificant (p-value = 0.566), suggesting that women and men are equally productive individually. A-level GPA strongly predicts performance: a one standard deviation improvement in GPA improves performance by 0.25 standard deviations. Finally, individual performance on the task is not systematically related to the student's field of study. A second regression shows no significant gender difference in how much subjects liked the task (p-value = 0.232).

The absence of productivity differences between female and male subjects is further corroborated by evidence from stage 2. After having met their potential partner in the audio chat room, subjects stated their belief regarding the productivity of the other subject when working individually on a task similar to the team task. Online Appendix Table A.12 shows that the other subject's gender does not significantly affect the belief subjects hold about the productivity of that person. We conclude that (in addition to women and men being equally productive at the task) the subjects believe that gender does not affect productivity.

Despite the absence of systematic gender differences in how productive subjects were individually or how much they liked the task, we cannot generally rule out

[27] The question read: "In your perception, how many of the other members of your group were women?" Hence, women had to distinguish between the *total* number of women (including themselves) and the number of women among their teammates. Males did not have to make this distinction.

[28] The recruiting of subjects for the individual task was identical to the team task. Subjects worked individually on the same task in the same online environment. The only difference was the absence of the audio chat (i.e., no interactions with other subjects in the session). Sessions ended after stage 1.

gender differences in how the subjects perceived the task. For instance, we cannot preclude that women and men stereotypically perceived (parts of) the team task to be inside or outside of their gender's domain. Using data from a laboratory experiment, Coffman (2014) shows that individuals collaborating in small groups are less willing to contribute ideas in areas that they consider to be outside of their gender's domain. Hence, unobserved gender differences in task perception could still contribute to systematic differences in the willingness to contribute to the team conversation. Interestingly, Coffman (2014) finds an overall gender difference in the willingness to contribute ideas that mirrors the lower number of words contributed by women in our experiment.[29]

4 Results

This section presents our main findings. We first discuss the evidence originating from stage 1. In cases where the team-level results are merely aggregating individual-level results, we comment on the team-level findings but relegate tables and figures to the Online Appendix.

4.1 Effects on Team Communication, Perceptions, and Performance

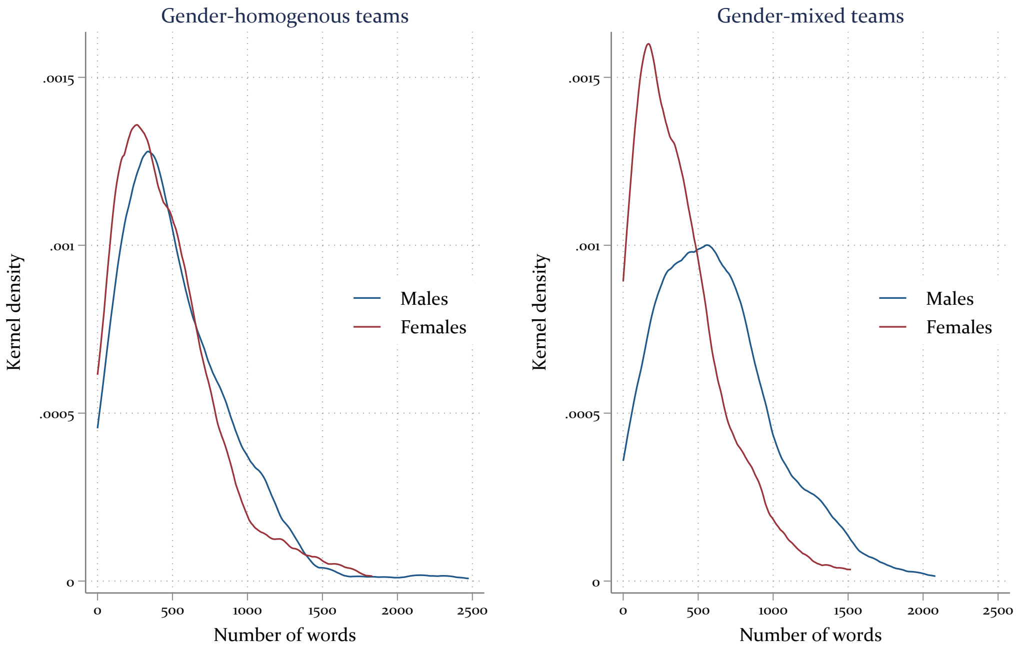
Quantity of Communication, Individual Level We begin by showing how the team gender composition affects the quantity of individual-level contributions to the team conversation. As a first step, Figure 1 presents individual-level kernel density plots for the number of words spoken. The figure shows a moderate shift to the right of the kernel density of men in all-male teams relative to the density of women in all-female teams. The difference between the respective densities in gender-mixed teams is much stronger. Comparing the densities between panels suggests that men tend to talk more in mixed teams relative to the counterfactual of working in a gender-homogenous team, whereas women adjust in the opposite direction.[30]

Estimation results using our primary quantitative outcomes are shown in Table 3. Columns (1) and (3) report the pre-specified regressions following equation (2). Both regressions show a consistent pattern, revealing quite dramatic gender differences in team communication. Relative to the mean of 519.4 words (38.6 turns) spoken by men in all-male teams, women in all-female teams speak 76.3 words (4.0 turns) less (β_1). Males in mixed teams speak 93.0 words (6.2 turns) more relative to men in all-male teams. The regressions also reveal a pronounced gender difference in how assignment

[29] In contrast to our study, the task used by Coffman (2014) is not gender-neutral in individual productivity.

[30] Figure B.5 in the Online Appendix shows kernel densities for the number of turns.

Figure 1: Total Number of Words, Individual Level



Notes: This figure shows kernel density plots for the number of words spoken at individual level, for subjects assigned to gender-homogenous ($N = 916$) and mixed teams ($N = 452$).

to a mixed rather than a gender-homogenous team affects communication (β_3). Equivalently, β_3 measures the difference in how female gender affects communication between mixed and gender-homogenous teams. Summing up β_2 and β_3 shows that women who were assigned to a mixed team speak 80.4 words (4.7 turns) less than women in all-female teams. Hence, whereas mixed teamwork makes men communicate more relative to gender-homogenous teamwork, for women the opposite is true. As a result, communication in mixed teams is heavily dominated by men. Taking the sum of β_1 and β_3 shows that in mixed teams, women on average speak about 250 words (15 turns) less than men. This implies that in mixed teams, men utter about 69 percent (50 percent) more words (turns) than women.

Some further observations from Table 3 are worth noting. First, accounting for multiple hypotheses testing leaves all our findings unchanged. Second, more cognitively skilled subjects speak significantly more, providing ex-post justification for our effort to ensure symmetry in the team-level composition of cognitive skills across teams of different gender compositions. On average, an improvement in the A-level GPA by one standard deviation makes a subject speak 69.5 words (3.6 turns) more, equivalent to 0.192 (0.155) standard deviations. Third, Columns (2) and (4) show that adding the Big 5 personality traits as further controls leaves all our main findings unchanged, but leads to a strong increase in the adjusted R^2. We conclude that

Table 3: Effects on the Quantity of Communication, Individual Level

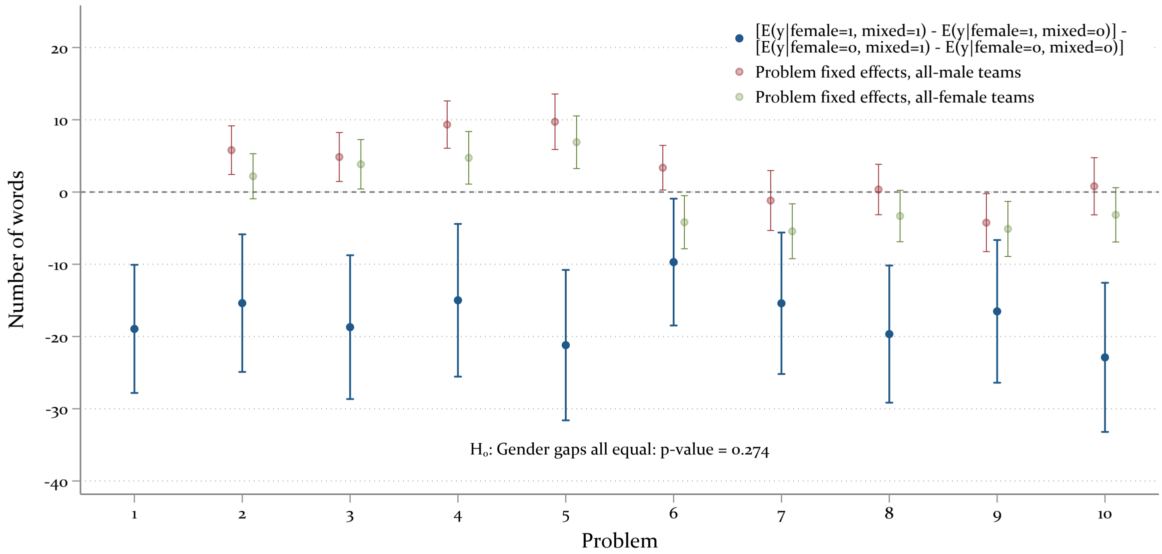
	#Words (1)	#Words (2)	#Turns (3)	#Turns (4)
Female (β_1)	-76.34***	-81.18***	-4.02**	-4.18**
	(23.49)	(24.25)	(1.77)	(1.76)
Mixed team (β_2)	93.03***	99.10***	6.16***	6.61***
	(28.74)	(28.07)	(2.12)	(2.01)
Female $\times$ Mixed team (β_3)	-173.39***	-182.98***	-10.87***	-11.31***
	(38.43)	(38.17)	(2.66)	(2.61)
A-level GPA	113.36***	116.77***	5.87***	6.46***
	(15.29)	(15.04)	(0.96)	(0.95)
Subject-level controls	Yes	Yes	Yes	Yes
Controls include Big 5	No	Yes	No	Yes
N. of obs.	1368	1281	1368	1281
Adj. R^2	0.100	0.207	0.085	0.204
Mean dep. var. all-male	519.4	517.0	38.6	38.3
$\beta_4 := \beta_1 + \beta_3$	-249.7	-264.2	-14.9	-15.5
$\beta_4 = 0$ (p-value)	0.000	0.000	0.000	0.000
$\beta_5 := \beta_2 + \beta_3$	-80.4	-83.9	-4.7	-4.7
$\beta_5 = 0$ (p-value)	0.001	0.001	0.009	0.008
$\beta_1 = 0$ (p-value MHT)	0.003	0.002	0.025	0.021
$\beta_2 = 0$ (p-value MHT)	0.005	0.002	0.006	0.003
$\beta_3 = 0$ (p-value MHT)	0.000	0.000	0.000	0.000

Notes: This table shows OLS regressions using as dependent variables the number of words and the number of turns at the individual level, respectively. Standard errors (clustered at team level) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT) follow Barsbai et al. (2020). Multiple testing is done across Columns (1) and (3) and Columns (2) and (4), respectively.

conditional on covariates including cognitive skills, personality traits are important drivers of communication behavior. Online Appendix Table A.13 reports the effects of the Big 5 personality traits and shows the robustness of the results to a specification without controls and a specification using log-transformed outcomes. For illustration purposes, Table A.14 in the Online Appendix shows regressions equivalent to those in Table 3 using total speaking time as an outcome. Furthermore, Online Appendix Figure B.6 shows that in 78% of all mixed teams, a male subject ranks first in terms of the number of words uttered.

A final observation from the subject-level communication data refers to how the observed differences in communication behavior evolve over time. As documented by Born et al. (2022), women tend to have lower self-confidence than men in team environments. It could be that in our setting, the gender gap in self-confidence is particularly pronounced at the beginning of the team interaction, but then becomes attenuated with the increasing familiarity of the team members with each other and the setting. Addressing this concern, Figure 2 demonstrates that the patterns in subjects' communication behavior are very stable over the 10 problems of the team

Figure 2: Gender Gap in Number of Words by Problem, Individual Level



Notes: This figure is derived from an OLS regression of equation (3). The figure displays problem-specific gender gaps $\hat{\theta}_p$ for $p = 1, \ldots, 10$ (blue dots), together with 95% confidence intervals. For comparison, the figure also displays $\hat{\beta}_p$ for $p = 2, \ldots, 10$ (problem fixed effects for males in all-male teams, red dots). The problem fixed effects for women in all-female teams (green dots) are derived from an equivalent regression that uses an indicator for males (plus corresponding interactions) instead of an indicator for females. The estimations use all $1386 \times 10 = 13860$ observations.

task. The figure uses subject-by-problem panel data and displays coefficients obtained from an OLS regression of equation (3). The blue dots show the estimated coefficients $\hat{\theta}_p$ for $p = 1, \ldots, 10$, together with 95% confidence intervals based on standard errors accounting for team-level clusters. Each $\hat{\theta}_p$ captures the problem-specific gender difference in how assignment to a mixed rather than a gender-homogenous team affects communication, thus providing a problem-specific disaggregation of $\hat{\beta}_3$ from Table 3, Column (1). For all 10 problems, $\hat{\theta}_p$ is negative and significant at least at the 5 percent level, and the hypothesis that all $\hat{\theta}_p$ are equal cannot be rejected (p-value $=$ 0.274).[31]

Quantity of Communication, Team Level Table 4 reports team-level regressions.[32]

Columns (1) and (2) report the pre-specified regressions using as outcomes the number of words and turns, respectively. In terms of the quantity of communication, all-male teams rank first, followed by mixed teams, while all-female teams are the least communicative. However, the mixed-team effect (β_1) is estimated imprecisely. As an exploratory analysis, the table reports in Column (3) a regression using as an outcome the number of words that are topically related to the problems the teams were working on. To construct the dependent variable, we collected from the problem sets

[31] For an equivalent analysis of turns, see Figure B.7 in the Online Appendix.
[32] Figure B.8 in the Online Appendix shows team-level kernel density plots for words and turns.

Table 4: Effects on the Quantity of Communication, Team Level

	#Words	#Turns	#Topic words
	(1)	(2)	(3)
Gender-mixed team (β_1)	-134.68	-5.89	-12.15**
	(86.36)	(6.72)	(4.70)
All-female team (β_2)	-297.51***	-16.58**	-20.24***
	(94.63)	(7.41)	(5.16)
N. of obs.	342	342	342
Mean dep. var. all-male	2077.7	154.5	127.3
Team-level controls	Yes	Yes	Yes
$\beta_1 = \beta_2$ (p-value)	0.079	0.131	0.093
$\beta_1 = 0$ (p-value MHT)	0.169	0.371	0.025
$\beta_2 = 0$ (p-value MHT)	0.009	0.070	0.000

Notes: This table shows OLS regressions at the team level. The dependent variables are the number of words spoken, the number of turns, and the number of words that are topically related to the team task, respectively. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT, six hypotheses included) follow Barsbai et al. (2020).

all words that were specific in the sense that it would be unlikely that the teams would frequently use these in a conversation unrelated to the problems (like "innovation capital", "investment", or "market share"). To account for references to the four possible solutions (labeled from *a* to *d*), we added to the list the expressions "A", "B", "C", and "D". We then derived the set of topic words by selecting from the list (separately for each problem set) the 10 most frequently used words. Even with this narrowly defined set, the topic words account for more than two thirds of mentions of all words listed, and about 7 percent of all words uttered.[33] Appendix Table A.15 displays the word lists used to define topic words and how frequently these were used.

Column (3) in Table 4 reveals that all-male teams use more words that are topically related to the team task than mixed and all-female teams. Table A.16 in the Online Appendix shows that the latter finding is robust to broader definitions of the set of topic words.

Team Performance Table 5 shows how the team gender composition affects team performance, measured by how many of the 10 problems a team solved.

We find that all-male teams outperform both gender-mixed and all-female teams, whereas the hypothesis of no difference between mixed and all-female teams cannot be rejected. On average, gender-mixed teams solve 0.4 problems (8.7 percent) less relative to all-male teams. All-female teams solve 0.55 problems less than all-male

[33] This is mainly due to the fact that references to the response options "A", "B", "C", and "D" alone make up almost 50 percent of all mentions of the listed words.

Table 5: Effects on Team Performance

	Number of problems solved
Gender-mixed team (β_1)	-0.402*
	(0.225)
All-female team (β_2)	-0.550**
	(0.254)
N. of obs.	342
Mean dep. var. all-male	4.61
Team-level controls	Yes
$\beta_1 = \beta_2$ (p-value)	0.529
$\beta_1 = 0$ (p-value MHT)	0.083
$\beta_2 = 0$ (p-value MHT)	0.062

Notes: This table shows an OLS regression using as dependent variable the number of problems solved at the team level. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT, two hypotheses included) follow Barsbai et al. (2020).

teams on average, implying a performance gap of 11.9 percent relative to all-male teams.[34]

While not the focus of our study, it is still of interest how teams perform relative to individuals. Teams can pool information and ideas, making them more productive than individuals. On the other hand, teams could be less productive because of free-riding and/or failure to coordinate the answers submitted by the individual team members (Grosse et al., 2011). Whereas free-riding incentives are common in team contexts, the coordination requirements were specific to the experimental design (aiming at inducing communication) and thus rather artificial. Appendix Table A.18 compares performance between teams and individuals working under the individual piece rate scheme. Overall, the impacts of free-riding and coordination failure cancel out the benefits of teamwork resulting from information pooling. Considering only teams that successfully coordinated (thus netting out the artificial aspects of the team task), we find that teams outperform individuals by 0.44 problems, or 10.2 percent.[35]

[34]As discussed in Section 2.3, our estimation sample comprises 31 teams with a team silent member. Since we did not foresee that subjects would stay silent, we did not specify in the pre-analysis plan how to treat those. If we exclude all teams with a silent member, the coefficients β_1 and β_2 remain almost unchanged and significant (p-values: 0.077 and 0.030). Regarding covariates, we stated in the pre-analysis plan that we would check if excluding the minimum and maximum of A-level GPA and age affects our findings. Doing so, we again obtain very similar coefficients (p-values: 0.078 and 0.058). Table A.17 in the Online Appendix reports additional regressions that split up the mixed-team coefficient by the different compositions in terms of gender by cognitive skills.

[35]The team task could have the property of a "maximum" production function. If true, the differences in team performance could be explained by best performers in the task being predominantly male. Using the data on individual performance, we do not find support for this notion: Out of the 149 men who worked on the task individually, 14 solved 7, 5 solved 8, and 1 solved 9 problems. Out of 147 women, 12 solved 7, 4 solved 8, and 1 solved 9 problems.

Table 6: Quantity of Communication and Team Performance

	Number of problems solved
#all words (β_1)	-0.001**
	(0.000)
#topic words (β_2)	0.015***
	(0.004)
N. of obs.	342
Mean dep. var.	4.35
Team-level controls	Yes

Notes: This table shows an OLS regression using as dependent variable the number of problems solved at the team level. The regression does not condition on team gender composition but uses as regressors of interest the overall number of words and the number of words that are topically related to the team task. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Quantity of Communication as Channel The next question we ask is whether the gender composition affects team performance through the quantity of communication, or whether this is an unlikely channel. We did not pre-register this analysis, which is therefore of an exploratory nature. The analysis starts from two facts that we have already established: First, all-male teams outperform mixed and all-female teams. Second, differences in team performance can only emerge through communication. This is because, by design, the teams could solve a given problem (and team members could earn a bonus for this problem) only if all team members chose the same (correct) answer. Teams, therefore, had to coordinate, and communication via the audio chat was the only available channel.

Differences in team performance could emerge through either the quantity or the quality of communication (or both).[36] By quality of communication, we mean the potential of an utterance to help a team find the correct answer to a given problem.[37] We first explore the quantity channel. Starting from the finding that all-male teams use more topic words than mixed and all-female teams, we study if the quantity of team communication correlates with team performance. Table 6 reports a team-level regression that does not condition on team gender composition but uses as main regressors the overall number of words spoken and the number of topic words. We find that conditional on the overall quantity of communication, teams that use more topic words perform better. A one-standard deviation increase in the number of topic words is associated with a shift in performance by 0.57 problems, or 0.34 standard deviations. Holding the number of topic words constant, teams that communicate

[36]Note that differences in free-riding would ultimately show up as differences either in the quantity or the quality of individual contributions to the team conversation.

[37]We also studied coordination as a potential quality dimension. Table A.19 shows that the team gender composition does not affect the ability of teams to coordinate, and that we obtain similar results as in Table 5 if we use only teams that manage to coordinate in all 10 problems.

Table 7: No Gender Gap in Share of Topic Words

	Share of topic words
Female (β_1)	0.001
	(0.002)
Mixed team (β_2)	0.000
	(0.002)
Female $\times$ Mixed team (β_3)	0.001
	(0.003)
A-level GPA	-0.001
	(0.001)
N. of obs.	1336
Mean dep. var. all-male	0.065
Subject-level controls	Yes
$\beta_1 + \beta_3 = 0$ (p-value)	0.538
$\beta_2 + \beta_3 = 0$ (p-value)	0.708

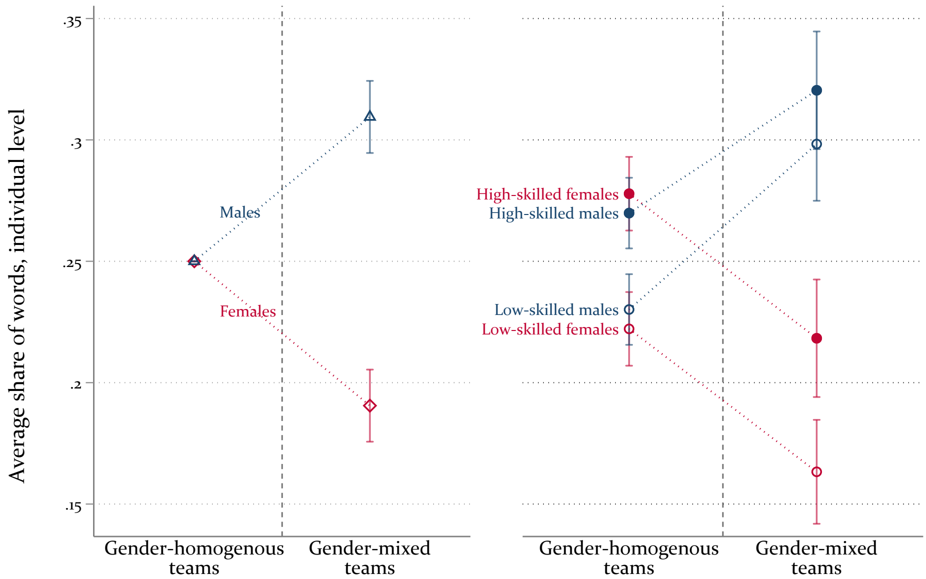
Notes: This table shows a subject-level OLS regression using as dependent variable the share of words in a subject's utterances that are topically related to the team task. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

more perform slightly worse, suggesting that teams that talk more about topics not directly related to the team task get distracted. Again, the analysis is robust to broader definitions of the set of topic words (see Online Appendix Table A.20).

We next explore the quality channel. Acknowledging that measuring the informational content of verbal communication is a challenge, we use the share of topic words in all words spoken as a proxy for quality. To test if this proxy has any power in explaining why all-male teams outperform mixed and all-female teams, we estimate equation (2), using as an outcome the share of topic words in all words uttered by subject i. Table 7 shows that utterances offered by women and men on average do not differ in the share of topic words. This holds irrespective of whether subjects work in a gender-homogenous or gender-mixed team. Using broader sets of topic words leads to very similar findings (Online Appendix Table A.21). We conclude that, to the extent that the share of topic words is a reasonable proxy for the quality of communication, performance differences between all-male teams and teams of alternative gender composition cannot be explained by differences in the quality of team communication.

Overall, our exploratory analysis on channels provides suggestive evidence that the gender composition affects team performance through the quantity of communication that is topically related to the team task. The quality of communication (proxied by the share of topic words) is unaffected by the team gender composition, and thus cannot be a channel through which the gender composition affects team performance.

Figure 3: Gender Gap in Team Communication: Share of Words



Notes: This figure uses the raw data and displays gender gaps in team communication by team gender composition and cognitive skills, together with 95% confidence intervals. The left panel shows shares in the total number of words at the team level spoken by women and men, separately for gender-homogenous and gender-mixed teams. The right panel differentiates between subjects of above-median ("high-skilled") and below-median ("low-skilled") cognitive skills in terms of A-level GPA. The sample consists of all 1386 subjects.

Determinants of Communication Behavior: Gender vs. Cognitive Skills Next, we analyze for further illustration the role of cognitive skills vs. gender as determinants of communication behavior. We did not pre-register this analysis, which is therefore of an exploratory nature.

Figure 3 visualizes word shares of women and men across teams of different gender compositions.[38] The left panel displays the average word shares women and men contributed to the team conversation. In gender-homogenous teams, these shares are mechanically equal to 25 percent. In gender-mixed teams, the average shares of words of women and men are 19.1 and 30.9 percent, respectively. This difference is highly significant (p-value < 0.01). The right panel splits the gender-specific means by skill level (above vs. below median). First, cognitive skills strongly predict communication shares in gender-homogenous teams. In both all-female and all-male teams, subjects of above-median skills talk significantly more than their teammates of below-median skills (both p-values < 0.01). Second, gender strongly dominates skills in predicting communication behavior in gender-mixed teams. On average, men of below-median skills contribute 29.8 percent of all words in mixed teams, only 2.2 percentage points less than men of above-median skills. This difference is not

[38] Figure B.9 in the Online Appendix shows similar patterns for the share of turns.

statistically different from zero (p-value = 0.245). Strikingly, men of below-median skills talk significantly more than women of above-median skills, whose word share is about 8 percentage points lower on average (p-value < 0.01). Hence, considering within-team shares of communication, gender-mixed teams have the striking feature of allowing men of below-median skills to elevate themselves to above-average positions while marginalizing women of above-median skills.[39] Women of below-median skills have even lower word shares and on average contribute only 16.3 percent to the team conversation. The difference relative to women of above-median skills is highly significant (p-value < 0.01). A further observation from the right panel of Figure 3 is that the difference in word shares between subjects of above-median and below-median skills is more pronounced among women than among men. However, neither in gender-homogenous (p-value = 0.455) nor in gender-mixed teams (p-value = 0.121) the respective difference is statistically different from zero.

Leadership is valuable in team settings (Englmaier et al., 2021). Despite the fact that the teams in our design had no formal leader, and there was no design element to initiate a discussion about leadership or make the teams choose a leader, Figure 3 links our work to the literature on individuals' aspirations to lead. This literature shows that attributions of leader emergence tend to be correlated with speaking time (Schmid Mast, 2002; MacLaren et al., 2020). The gender gaps in speaking time in our setting are thus in line with evidence that women are less willing than men to strive for team leadership positions (Alan et al., 2020; Born et al., 2022). Our design contrasts all-female and all-male teams with gender-balanced teams, preventing us to study teams with a majority of either women or men. Evidence supporting the idea that teams with a majority of women provide a more favorable environment for women is provided by Karpowitz et al. (2023), who show that increasing the share of women or assigning a female leader reduces discrimination against women.

Distribution of Team Communication Table A.22 in the Online Appendix uses equation (1) to analyze how the team gender composition affects the distribution of communication at team level, measured by Herfindahl-Hirschmann indices of words and turns, respectively. We find that β_1 and β_2 are both insignificant, suggesting that the degree of inequality in communication in mixed and all-female teams does not differ significantly from all-male teams. However, s$H_0 : \beta_1 = \beta_2$ can be rejected at conventional levels, implying that communication in mixed teams is distributed more unequally relative to all-female teams.

[39]The latter finding relates our work to Shan (2022), who also studies small groups and finds that women in a minority position interact less with peers and have lower self-confidence.

Table 8: Effects on Sentiment, Individual Level

	Positive (1)	Negative (2)
Female (β_1)	0.260***	-0.064***
	(0.014)	(0.013)
Mixed team (β_2)	-0.002	-0.000
	(0.017)	(0.015)
Female × Mixed team (β_3)	-0.035*	0.037**
	(0.021)	(0.018)
N. of obs.	1336	1336
Mean dep. var. all-male	0.26	0.28
Subject-level controls	Yes	Yes
$\beta_4 := \beta_1 + \beta_3$	0.225	-0.027
$\beta_4 = 0$ (p-value)	0.000	0.021
$\beta_5 := \beta_2 + \beta_3$	-0.037	0.037
$\beta_5 = 0$ (p-value)	0.008	0.009
$\beta_1 = 0$ (p-value MHT)	0.000	0.000
$\beta_2 = 0$ (p-value MHT)	0.977	0.996
$\beta_3 = 0$ (p-value MHT)	0.202	0.115

Notes: This table shows OLS regressions using as dependent variables measures of individual sentiment of team communication captured by vocal features. Positive (negative) sentiment captures vocal features indicating happiness (sadness). Standard errors (clustered at team level) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT, six hypotheses included) follow Barsbai et al. (2020).

Sentiment of Communication Table 8 reports individual-level OLS regressions of equation (2) using vocal measures of sentiment as dependent variables.[40] The approach regards voice as a digital signal and focuses on the physical information revealing the speaker's emotions. We split the vocal samples by the speaker's gender and separately trained a female and a male model. Using the algorithm, we construct three turn-specific weights: positive (capturing happiness), negative (sadness), and neutral. If a turn was spoken by a woman (man), we used the female (male) model to construct the weights. We obtain measures of positive, negative, and neutral sentiment by taking averages over turns, weighted by the turns' length.[41]

The estimates of β_1 indicate that utterances by women working in all-female teams carry more positive sentiment (vocal features indicating happiness) and less negative sentiment (vocal features indicating sadness) relative to men in all-male teams. We would like to caution, however, that we obtain our sentiment measures from two separate gender-specific models trained to classify emotions. As a result, different

[40]In the pre-analysis plan, we committed to capture the polarity of communication by a lexical score. As we demonstrate in the Online Appendix, Section D, the lexical sentiment score turned out to be dominated by the teams' usage of words likely triggered by the single-choice design of the team task. We, therefore, decided to rely on measures of team sentiment based on vocal features.

[41]For further details, see Online Appendix Section C. We pre-specified to treat the polarity of team sentiment as a primary outcome only at the team level. For completeness, we also report the evidence on sentiment at the individual level.

estimates of β_1 could at least partly reflect differences between models rather than true emotions. By contrast, β_2 and β_3 are identified by differences in emotions within gender and thus cannot be driven by differences between models. Interestingly, the estimates of β_2 indicate that the vocal sentiment of men is independent of whether they work in all-male or mixed teams. Hence, there is no evidence that men establish dominance in mixed teams via a particular communication style towards women. By contrast, the sentiment of women in mixed teams is less positive and more negative relative to all-female teams ($H_0 : \beta_2 + \beta_3 = 0$, p-values < 0.01), possibly a response to men dominating the team conversation quantitatively.

Table A.23 in the Online Appendix presents team-level regressions for the sentiment of communication following equation (1). We find a clear ranking in the extent to which positive emotions (happiness) characterize the team communication: all-female teams communicate more positively than gender-mixed teams, and gender-mixed teams more positively than all-male teams. Negative emotions (sadness) are less prevalent in all-female teams relative to mixed and all-male teams.[42]

Perceptions About Team Interaction Table 9 reports estimations of equation (2) using perceptions about team interaction as dependent variables. Our survey-based measures of perceptions capture the positivity of team communication, the cooperativeness of team communication, and the likeability of the team task. These perceptions were elicited individually in the survey at the end of stage 1 using 5-point Likert scales.[43] In regressions using the three perceptions separately, we find little evidence for systematic effects of team gender composition. Column (4) complements the analysis using a z-score index of perceptions following Kling et al. (2007).[44] There is no evidence that women and men from gender-homogenous teams differ in their perceptions of team interaction. Males from mixed teams tend to have higher index values than men from all-male teams, but the difference is not significantly different from zero. Women working in mixed teams tend to have lower index values than women in all-female teams, but the difference is also not significantly different from zero. Whereas the point estimates of β_1 to β_3 are all insignificant, $H_0 : \beta_1 + \beta_3 = 0$ can be rejected (p-value $= 0.036$), suggesting that in mixed teams, women perceive the team interaction to be worse relative to men. Table A.24 in the Online Appendix

[42]We would like to reiterate that these differences could partly be due to the use of gender-specific models when classifying emotions.

[43]Subjects were asked to what extent they agree to the following: "The communication in my group was characterized by a positive tone" (positivity), "The communication in my group was cooperative" (cooperativeness), and "Working on the problems together was fun" (likeability). Higher values indicated stronger agreement. Team-level measures are averages over subject-level values in a team.

[44]To construct the index, we standardize each outcome into a z-score by subtracting the mean among men in all-male teams and dividing by the respective standard deviation. We then average all the z-scores and again standardize to men from all-male teams.

Table 9: Effects on Perceived Team Interaction, Individual Level

	Positivity (1)	Cooperativeness (2)	Likeability (3)	Perception index (4)
Female (β_1)	0.001	-0.007	-0.082	-0.042
	(0.049)	(0.045)	(0.074)	(0.082)
Mixed team (β_2)	0.033	-0.001	0.133*	0.081
	(0.053)	(0.050)	(0.078)	(0.078)
Female $\times$ Mixed team (β_3)	-0.100	-0.032	-0.177*	-0.166
	(0.080)	(0.082)	(0.103)	(0.129)
N. of obs.	1358	1357	1362	1356
Mean dep. var. all-male	4.66	4.66	4.06	0.03
Subject-level controls	Yes	Yes	Yes	Yes
$\beta_4 := \beta_1 + \beta_3$	-0.099	-0.038	-0.259	-0.208
$\beta_4 = 0$ (p-value)	0.123	0.578	0.000	0.036
$\beta_5 := \beta_2 + \beta_3$	-0.068	-0.033	-0.044	-0.085
$\beta_5 = 0$ (p-value)	0.325	0.615	0.623	0.451
$\beta_1 = 0$ (p-value MHT)	0.984	0.998	0.785	0.625
$\beta_2 = 0$ (p-value MHT)	0.959	0.999	0.431	0.494
$\beta_3 = 0$ (p-value MHT)	0.727	0.990	0.436	0.430

Notes: This table shows OLS regressions using as dependent variables different outcomes measuring individual perceptions about team interaction. Perceived positivity, cooperativeness, and likeability of the team task are all measured using a 5-point Likert scale. The perception index is constructed by aggregating standardized perceptions in all three dimensions (Kling et al., 2007). Standard errors (clustered at team level) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT, nine hypotheses across Columns (1) to (3) and three hypotheses in Column (4)) follow Barsbai et al. (2020).

reports team-level regressions of perceptions. None of the coefficients is estimated to be significantly different from zero.

Given the dominance of men in mixed teams and the effects of team gender composition on team sentiment, it is surprising that we see little corresponding effects on perceptions of team interaction. We can think of two possible explanations. First, it is possible that the non-incentivized elicitation of perceptions results in measurement error and imprecise coefficient estimates. A second possibility is that the subjects are used to communication following the patterns analyzed before, including male dominance in gender-mixed settings, and simply perceive these patterns as normal, or in line with expectations. The fact that we identify the effects of team gender composition from between-subject variation could then explain the absence of strong differences in measured perceptions.[45]

[45]Tables A.25 and A.26 in the Online Appendix report additional regressions using perceptions that we pre-specified as secondary outcomes (whether the team's communication was sufficient, whether it was symmetric, and whether subjects let each other finish). The only significant effects we find indicate that women in all-female teams perceived the team communication to be more symmetric relative to men in all-male teams.

Evidence from Stage 1: Discussion The analysis of stage 1 of the experiment has delivered the following main insights. First, all-male teams communicate more than mixed and all-female teams. These differences are more pronounced if we focus on topic words rather than all words spoken. Second, all-male teams outperform both gender-mixed and all-female teams, and an exploratory analysis suggests that team performance is driven by the usage of topic words. Third, in mixed teams, men dominate the team communication quantitatively.

A crucial question refers to the channels through which variation in the team gender composition impacts the subjects' communication behavior. Two fundamentally different explanations for why men are more talkative than women (in particular in mixed teams) stand out. First, it could be that men have a stronger preference for dominance and put through this preference by means of an aggressive communication style. Second, the observed differences in communication behavior could be due to gender differences in self-confidence and established gender roles that are so deeply ingrained in the subjects' beliefs about adequate social behavior that male dominance emerges without any battle of the sexes about communication shares.

Regarding the first hypothesis, an emerging literature documents that men often establish dominance over women through hostility and interruptions (Jacobi and Schweers, 2017; Dupas et al., 2021; Miller and Sutherland, 2023). Overall, we find little support for the notion that something similar happened in our setting. The evidence on individual sentiment (Table 8) suggests that the emotions conveyed in the voices of men interacting in mixed teams are no different from those in all-male teams. Furthermore, there is little evidence that male dominance is established through interruptions. Figure B.10 in the Online Appendix documents that in mixed teams, there is no gender difference in active interruptions, and women are only slightly more likely to be interrupted by others. Figure B.11 shows that in mixed teams, the share of interruptions of female speakers caused by men is larger than the share of interruptions of male speakers caused by women, but this does not change the fact that overall, women and men face similar chances of being interrupted.

Regarding the second hypothesis, the literature has documented that women have lower self-confidence (Kling et al., 1999; Croson and Gneezy, 2009), lower social confidence (Alan et al., 2020), and downgrade their self-assessment when observed by others more strongly than men (Ludwig et al., 2017). In line with this evidence, in the sample of subjects who worked on the team task individually, we find that men systematically overestimate their own performance, whereas women do not (results available upon request). Similarly, Online Appendix Table A.27 shows that women are significantly less optimistic than men regarding their own performance under an individual piece rate in a possible second-stage task. Furthermore, Appendix Table

A.28 shows that the incidence of phrases that indicate uncertainty is higher among women, but does not depend on whether or not subjects work in a mixed team.

Differences in self-confidence could explain why men are more willing than women to actively participate in the team conversation. In a team setting where contributing to the team's success required speaking in front of others, existing differences in self-confidence could be amplified through differences in social confidence, defined as the willingness to perform a task in public. Gender differences in self-confidence and social confidence could also translate into differences in the likelihood to contradict other team members and thereby override wrong solutions (Isaksson, 2019; Radbruch and Schiprowski, 2023; Guo and Recalde, 2023). The fact that women in mixed teams talk less than women in all-female teams is in line with Born et al. (2022), showing that women are less willing to lead male-majority teams due to a negative effect on their confidence. However, our data do not allow us to pin down the channel through which this effect works. Interestingly, the literature on classroom interaction has shown that children get accustomed to boys dominating the group communication in mixed-gender settings early on (for reviews, see Kelly 1988 and Aukrust 2008). This could explain why in our setting, we do not find significant traces of women and men competing for speaking time when interacting in gender-mixed teams.

Considering both hypotheses, we believe that the gender gaps in communication likely reflect gender differences in self- and social confidence and existing gender roles in accordance with these differences. Having discussed the evidence from stage 1, we now turn to the results on beliefs and preferences for teamwork from stage 2.

4.2 Effects on Preferences for Teamwork and Beliefs

Beliefs about Productivity and Communication We measure beliefs about own productivity, potential partner's productivity, and team productivity at a possible further task in stage 2 plus beliefs about the team interaction (positivity of team communication, cooperativeness of team communication, and likeability of the team task). To elicit productivity beliefs, we asked the subjects to imagine a task similar to the one in the first stage comprising 20 problems. All productivity beliefs take integer values between 1 and 20, depending on the subject's stated belief about how many problems she (the potential partner, the team) would solve. Beliefs about the social interaction with the potential partner in a possible further team task were measured in the same way as perceptions in stage 1 (5-point Likert scales).[46]

[46]For example, we elicited own-productivity beliefs as follows: "What do you think: If you were working on the task alone, how many of the 20 problems would you answer correctly?" For positivity beliefs, we asked for agreement (5-point Likert scale) with the statement: "The communication with the other person would be characterized by a positive tone." See the screenshots in the Online Appendix, Section E, for the wording of the other questions.

We report the pre-registered regressions studying beliefs regarding productivity and communication in a possible further team task in the Online Appendix and briefly summarize the findings here. In line with evidence on male overconfidence in comparable settings (e.g. Gneezy et al., 2003; Croson and Gneezy, 2009), Table A.27 demonstrates that, irrespective of the team gender composition in stage 1, women have less optimistic beliefs than men regarding their own performance in a possible second-stage task.[47] Women are also less optimistic regarding team productivity. By contrast, there are no significant effects on subjects' beliefs regarding their partner's productivity. The finding that productivity beliefs are unrelated to whether or not subjects were exposed to mixed teamwork in stage 1 is confirmed in Table A.29, reporting estimations that condition on the potential partner's gender.

Table A.30 presents estimation results for beliefs about communication. In addition to the pre-registered regressions, we study a z-score index over beliefs regarding positivity, cooperativeness, and the likeability of the team task in a possible team interaction with the potential partner. The results indicate that men hold more positive beliefs if they were assigned to a mixed team in stage 1. Table A.31 complements this evidence by estimations that also condition on the potential partner's gender. The estimations using the belief index as an outcome show that women who were assigned to an all-female team in the first stage hold more positive beliefs about the interaction with the potential partner if the partner is female. For women who were assigned to a gender-mixed team in stage 1, no such effect is present. For men, none of the coefficients is significant.

Overall, there is little evidence that past exposure to mixed teamwork affects subjects' productivity beliefs in significant ways. Similarly, beliefs about communication in further teamwork seem largely unaffected by past exposure to mixed teamwork. The latter finding is in line with the insight from Table 9 that the team gender composition had no systematic impact on subjects' perceptions of team interaction.

Preferences for Teamwork Table 10 presents estimation results for equation (2) with and without an interaction effect between indicators for mixed teams and female gender. The dependent variable is the indicator for subjects who stated a preference for teamwork with the potential partner over individual work in a possible second-stage task. None of the estimated coefficients is significantly different from zero in itself. However, the estimate of $\beta_5 := \beta_2 + \beta_3$ in Column (2) is negative, and the hypothesis $\beta_5 = 0$ can be rejected (p-value $= 0.089$), indicating that women who were assigned to

[47]Since the majority of subjects did not work on the second-stage task, we do not have any measure of overconfidence regarding individual performance for our experimental sample. However, in the sample of subjects who worked on the first-stage task individually (see Section 3.5 for details), we find that men are significantly more overconfident than women.

Table 10: Preferences: Past Exposure to Mixed Teamwork

	= 1 if subject prefers teamwork	
	(1)	(2)
Female (β_1)	-0.027	-0.002
	(0.031)	(0.036)
Mixed team (β_2)	-0.037	-0.000
	(0.031)	(0.043)
Female $\times$ Mixed team (β_3)		-0.076
		(0.062)
N. of obs.	731	731
Mean dep. var. all-male	0.81	0.81
Subject-level controls	Yes	Yes
$\beta_4 := \beta_1 + \beta_3$		-0.077
$\beta_4 = 0$ (p-value)		0.149
$\beta_5 := \beta_2 + \beta_3$		-0.076
$\beta_5 = 0$ (p-value)		0.089
$\beta_1 = 0$ (p-value MHT)	0.398	0.999
$\beta_2 = 0$ (p-value MHT)	0.407	0.999
$\beta_3 = 0$ (p-value MHT)		0.494

Notes: This table shows OLS regressions using as dependent variable an indicator for subjects who indicate that they prefer to work in a team with the potential partner (rather than work individually) on a possible further task. Standard errors (in parentheses) account for clusters comprising all subjects from first-stage teams used in the cross-wise random assignment to pairs of potential partners. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT, two hypotheses included in Column (2) and three hypotheses in Column (3)) follow Barsbai et al. (2020).

a mixed team in stage 1 have a lower preference for teamwork relative to women who were assigned to an all-female team.

Further analyzing preferences for teamwork, Table 11 presents estimates of equation (4), separately for female and male subjects. Column (1) shows that, first, the potential partner's gender does not affect preferences for teamwork for women who worked in all-female teams in stage 1. Second, the reduced tendency to opt for teamwork after exposure to mixed teamwork in stage 1 observed in Table 10 turns out to be specific to those women whose potential partner in stage 2 is male. In contrast, for women whose potential partner in stage 2 is female, past exposure to mixed teamwork does not significantly affect preferences for further teamwork. Column (2) shows that, similarly to women, the potential partner's gender does not affect men's preferences after exposure to gender-homogenous teamwork in stage 1. Interestingly, after exposure to mixed teamwork, men respond differently than women to their potential partner's gender. After having collaborated with a gender-mixed team in stage 1, men are significantly more likely to prefer teamwork over individual work if their potential partner is female.

In summary, Table 11 delivers at least suggestive evidence that past exposure to gender-mixed teamwork makes women *more reluctant* to engage in teamwork with

Table 11: Preferences: Past Exposure and Partner's Gender

	= 1 if subject prefers teamwork	
	Females	Males
	(1)	(2)
Female partner 2nd stage (β_1)	0.031	-0.006
	(0.059)	(0.046)
Mixed team 1st stage (β_2)	-0.111*	-0.090
	(0.066)	(0.065)
Female partner 2nd stage × Mixed team 1st stage (β_3)	0.058	0.210**
	(0.099)	(0.087)
N. of obs.	351	380
Mean dep. var. gender-homogenous teams	0.80	0.81
Subject-level controls	Yes	Yes
$\beta_4 := \beta_1 + \beta_3$	0.089	0.204
$\beta_4 = 0$ (p-value)	0.250	0.005
$\beta_5 := \beta_2 + \beta_3$	-0.053	0.120
$\beta_5 = 0$ (p-value)	0.435	0.025
$\beta_1 = 0$ (p-value MHT)	0.846	0.898
$\beta_2 = 0$ (p-value MHT)	0.367	0.497
$\beta_3 = 0$ (p-value MHT)	0.899	0.087

Notes: This table shows OLS regressions using as dependent variable an indicator for subjects who indicate that they prefer to work in a team with the potential partner (rather than work individually) on a possible further task. Standard errors (in parentheses) account for clusters comprising all subjects from first-stage teams used in the cross-wise random assignment to pairs of potential partners. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT, six hypotheses included) follow Barsbai et al. (2020).

men. In contrast, men are *more willing* to engage in teamwork with women after being exposed to gender-mixed teamwork. Our design does not allow us to disentangle the different channels through which these opposite effects might work. One possible interpretation is that women respond to male dominance in gender-mixed teamwork by a reduced willingness to collaborate with men, whereas the experience of dominating a gender-mixed team's communication further increases the preference of men for teamwork involving women.

5 Conclusion

Using an online experiment, we study how the team gender composition affects team communication, team performance, and preferences for further teamwork. Regarding the quantity of team communication, we demonstrate that all-male teams communicate more than mixed and all-female teams. These differences are more pronounced if we focus on words that are topically related to the team task rather than all words spoken. The gender gap in communication is largest in mixed teams, where men heavily dominate the team communication quantitatively. Analyzing the role of gender and cognitive skills for mixed-team communication, we find that high-skilled men talk the most, followed by low-skilled men. High-skilled women talk significantly less than low-skilled men and significantly more than low-skilled women. Regarding team performance, all-male teams outperform both gender-mixed and all-female teams, and an exploratory analysis suggests that team performance is driven by the usage of topic words. Exploring effects on attitudes, we find suggestive evidence that past exposure to gender-mixed teamwork makes women less willing to engage in gender-mixed teams, while for men the opposite is true.

Our findings carry a number of important implications. For instance, our results suggest that the gender composition impacts the amount of information exchanged in teams and show that all-male teams tend to communicate more actively. This may help to explain why part of the literature (including our study) finds that all-male teams outperform mixed and all-female teams. To the extent that gender-specific communication behavior is socially acquired, our findings call for more research on how to ensure that starting from early childhood, the voices of women are properly heard. Based on research suggesting that speaking time correlates with leadership aspirations, another implication of our study is that in small-stakes environments, women working in gender-mixed teams are less likely to collect leadership experience relative to women working in all-female teams. To the extent that past leadership experience positively affects subjects' willingness to lead and the quality of leadership they provide, a lack of female leadership experience in small-stakes environments may help to explain the sizeable gender gaps in leadership observed in high-stakes environments. It remains to be studied whether, in small-stakes environments, gender-homogenous teams can be superior to mixed teams in effectively supporting women in building up leadership experience.

Finally, our evidence also suggests that exposure to gender-mixed teams negatively affects women's willingness to engage in gender-mixed teamwork. Policies aiming at integrating women into traditionally male-dominated domains may thus be subject to two limiting factors. First, in the absence of effective countermeasures, women in gender-mixed teams are likely to be dominated by men communication-wise and may rarely advance to leadership positions. Second, post-integration, women may be less open to gender-mixed teamwork in similar settings relative to the counterfactual of no integration. It remains an open question to what extent communication behavior in teams is malleable, and which type of intervention aiming at more gender-balanced communication in mixed teams is most effective.

References

ADAMS, R. B., J. DE HAAN, S. TERJESEN, AND H. VAN EES (2015): "Board Diversity: Moving the Field Forward," *Corporate Governance: An International Review*, 23, 77–82.

ADAMS, R. B. AND D. FERREIRA (2009): "Women in the Boardroom and Their Impact on Governance and Performance," *Journal of Financial Economics*, 94, 291–309.

AHERN, K. R. AND A. K. DITTMAR (2012): "The Changing of the Boards: The Impact on Firm Valuation of Mandated Female Board Representation," *Quarterly Journal of Economics*, 127, 137–197.

AL-UBAYDLI, O. AND J. A. LIST (2013): "On the Generalizability of Experimental Results in Economics: With a Response to Camerer," NBER Working Papers No. 19666.

ALAN, S., S. ERTAC, E. KUBILAY, AND G. LORANTH (2020): "Understanding Gender Differences in Leadership," *Economic Journal*, 130, 263–289.

ALLIANCE FOR BOARD DIVERSITY AND DELOITTE (2021): "Missing Pieces Report: The Board Diversity Census of Women and Minorities on Fortune 500 Boards, 6th Edition," Released June 8, 2021, on `https://www2.deloitte.com/content/dam/Deloitte/us/Documents/us-missing-pieces-fortune-500-board-diversity-study-sixth-edition.pdf`.

APESTEGUIA, J., G. AZMAT, AND N. IRIBERRI (2012): "The Impact of Gender Composition on Team Performance and Decision Making: Evidence from the Field," *Management Science*, 58, 78–93.

ARBAK, E. AND M.-C. VILLEVAL (2013): "Voluntary Leadership: Motivation and Influence," *Social Choice and Welfare*, 40, 635–662.

AUKRUST, V. G. (2008): "Boys' and Girls' Conversational Participation Across Four Grade Levels in Norwegian Classrooms: Taking the Floor or Being Given the Floor?" *Gender and Education*, 20, 237–252.

AZMAT, G. AND B. PETRONGOLO (2014): "Gender and the Labor Market: What Have We Learned from Field and Lab Experiments?" *Labour Economics*, 30, 32–40.

BANDIERA, O., I. BARANKAY, AND I. RASUL (2013): "Team Incentives: Evidence from a Firm Level Experiment," *Journal of the European Economic Association*, 11, 1079–1114.

BARSBAI, T., V. LICUANAN, A. STEINMAYR, E. TIONGSON, AND D. YANG (2020): "Information and the Acquisition of Social Network Connections," NBER Working Paper No. 27346.

Bertrand, M. and K. F. Hallock (2001): "The Gender Gap in Top Corporate Jobs," *Industrial and Labor Relations Review*, 55, 3–21.

Blau, F. and L. Kahn (2017): "The Gender Wage Gap: Extent, Trends, and Explanations," *Journal of Economic Literature*, 55, 789–865.

Borghans, L., B. H. Golsteyn, J. J. Heckman, and J. E. Humphries (2016): "What Grades and Achievement Tests Measure," *Proceedings of the National Academy of Sciences*, 113, 13354–13359.

Born, A., E. Ranehill, and A. Sandberg (2022): "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" *Review of Economics and Statistics*, 104, 259–275.

Burkhardt, F., A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, et al. (2005): "A Database of German Emotional Speech," *Interspeech*, 5, 1517–1520.

Buser, T. and H. Yuan (2023): "Public Speaking Aversion," *Management Science*, 69, 2746–2760.

Campbell, K. and A. Mínguez-Vera (2008): "Gender Diversity in the Boardroom and Firm Financial Performance," *Journal of Business Ethics*, 83, 435–451.

Carter, A. J., A. Croft, D. Lukas, and G. M. Sandstrom (2018): "Women's Visibility in Academic Seminars: Women Ask Fewer Questions than Men," *PloS ONE*, 13, e0202743.

Center for American Women and Politics (2022): "Women Appointed to Presidential Cabinets," Released March 16, 2022, on `https://cawp.rutgers.edu/sites/default/files/resources/womenapptdtoprescabinets.pdf`.

Chapple, L. and J. E. Humphrey (2014): "Does Board Gender Diversity Have a Financial Impact? Evidence Using Stock Portfolio Performance," *Journal of Business Ethics*, 122, 709–723.

Charness, G., D. J. Cooper, and Z. Grossman (2020): "Silence is Golden: Team Problem Solving and Communication Costs," *Experimental Economics*, 23, 668–693.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree – An Open-Source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

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

Congressional Research Service (2022): "Women in Congress, 1917-2022: Service Dates and Committee Assignments by Member, and Lists by State and Congress," Released March 3, 2022, on https://sgp.fas.org/crs/misc/RL30261.pdf.

Cooper, D. J., K. Saral, and M. C. Villeval (2021): "Why Join a Team?" *Management Science*, 67, 6980–6997.

Croson, R. and U. Gneezy (2009): "Gender Differences in Preferences," *Journal of Economic Literature*, 47, 448–74.

Dahl, G. B., A. Kotsadam, and D.-O. Rooth (2021): "Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams," *Quarterly Journal of Economics*, 136, 987–1030.

Davenport, J. R. A., M. Fouesneau, E. Grand, A. Hagen, K. Poppenhaeger, and L. L. Watkins (2014): "Studying Gender in Conference Talks - Data from the 223rd Meeting of the American Astronomical Society," *Physics and Society, arXiv Preprint: 1403.3091*.

De Paola, M., R. Lombardo, V. Pupo, and V. Scoppa (2021): "Do Women Shy Away from Public Speaking? A Field Experiment," *Labour Economics*, 70, 102001.

Deming, D. J. (2017): "The Growing Importance of Social Skills in the Labor Market," *Quarterly Journal of Economics*, 132, 1593–1640.

Dupas, P., A. S. Modestino, M. Niederle, J. Wolfers, and The Seminar Dynamics Collective (2021): "Gender and the Dynamics of Economics Seminars," *NBER Working Paper No. 28494*.

Edin, P.-A., P. Fredriksson, M. Nybom, and B. Öckert (2022): "The Rising Return to Noncognitive Skill," *American Economic Journal: Applied Economics*, 14, 78–100.

Englmaier, F., S. Grimm, D. Grothe, D. Schindler, and S. Schudy (2021): "The Value of Leadership: Evidence from a Large-Scale Field Experiment," CESifo Working Paper No. 9273.

Englmaier, F., S. Grimm, D. Schindler, and S. Schudy (2023): "The Effect of Incentives in Non-Routine Analytical Team Tasks," *Journal of Political Economy*, forthcoming.

Ertac, S. and M. Y. Gurdal (2012): "Deciding to Decide: Gender, Leadership and Risk-Taking in Groups," *Journal of Economic Behavior and Organization*, 83, 24–30.

Gerlitz, J.-Y. and J. Schupp (2005): "Zur Erhebung der Big-Five-Basierten Persönlichkeitsmerkmale im SOEP," *DIW Research Notes, 4, 2005*.

Gneezy, U., M. Niederle, and A. Rustichini (2003): "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, 118, 1049–1074.

Graetz, G. and A. Karimi (2022): "Gender Gap Variation Across Assessment Types: Explanations and Implications," *Economics of Education Review*, 91, 102313.

Grosse, S., L. Putterman, and B. Rockenbach (2011): "Monitoring in Teams: Using Laboratory Experiments to Study a Theory of the Firm," *Journal of the European Economic Association*, 9, 785–816.

Guo, J. and M. P. Recalde (2023): "Overriding in Teams: The Role of Beliefs, Social Image, and Gender," *Management Science*, 69, 2239.2262.

Gächter, S., C. Starmer, and F. Tufano (2023): "Measuring "Group Cohesion" to Reveal the Power of Social Relationships in Team Production," *Review of Economics and Statistics*, forthcoming.

Haegele, I. (2022): "The Broken Rung: Gender and the Leadership Gap," Mimeo, University of Munich.

Hamilton, B. H., J. A. Nickerson, and H. Owan (2003): "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy*, 111, 465–497.

——— (2012): "Diversity and Productivity in Production Teams," in *Advances in the Economic Analysis of Participatory and Labor-Managed Firms*, Emerald Group Publishing Limited, 99–138.

Hinsley, A., W. J. Sutherland, and A. Johnston (2017): "Men Ask More Questions than Women at a Scientific Conference," *PloS ONE*, 12, e0185534.

Hjort, J. (2014): "Ethnic Divisions and Production in Firms," *Quarterly Journal of Economics*, 129, 1899–1946.

Hoogendoorn, S., H. Oosterbeek, and M. Van Praag (2013): "The Impact of Gender Diversity on the Performance of Business Teams: Evidence from a Field Experiment," *Management Science*, 59, 1514–1528.

Hoogendoorn, S. and M. Van Praag (2012): "Ethnic Diversity and Team Performance: A Field Experiment," *IZA Discussion Paper No. 6731*.

Hu, A. and S. Ma (2021): "Persuading Investors: A Video-Based Study," NBER Working Paper No. 29048.

Isaksson, S. (2019): "It Takes Two: Gender Differences in Group Work," Mimeo, University of Stockholm.

Ivanova-Stenzel, R. and D. Kübler (2011): "Gender Differences in Team Work and Team Competition," *Journal of Economic Psychology*, 32, 797–808.

Jacobi, T. and D. Schweers (2017): "Justice, Interrupted: The Effect of Gender, Ideology, and Seniority at Supreme Court Oral Arguments," *Virginia Law Review*, 103, 1379–1496.

Karpowitz, C. F., S. D. O'Connell, J. Preece, and O. Stoddard (2023): "Strength in Numbers: Gender Composition, Leadership, and Women's Influence in Teams," IZA Discussion Papers No. 16625.

Kelly, A. (1988): "Gender Differences in Teacher-Pupil Interactions: A Meta-Analytic Review," *Research in Education*, 39, 1–23.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75, 83–119.

Kling, K. C., J. S. Hyde, C. J. Showers, and B. N. Buswell (1999): "Gender Differences in Self-Esteem: A Meta-Analysis." *Psychological Bulletin*, 125, 470–500.

Kuhn, P. and M. C. Villeval (2015): "Are Women More Attracted to Co-Operation than Men?" *Economic Journal*, 125, 115–140.

Lamiraud, K. and R. Vranceanu (2018): "Group Gender Composition and Economic Decision-Making: Evidence from the Kallystée Business Game," *Journal of Economic Behavior & Organization*, 145, 294–305.

Lazear, E. and K. L. Shaw (2007): "Personnel Economics: The Economist's View of Human Resources," *Journal of Economic Perspectives*, 21, 91–114.

List, J. A., A. M. Shaikh, and Y. Xu (2019): "Multiple Hypothesis Testing in Experimental Economics," *Experimental Economics*, 22, 773–793.

Ludwig, S., G. Fellner-Röhling, and C. Thoma (2017): "Do Women Have More Shame than Men? An Experiment on Self-Assessment and the Shame of Overestimating Oneself," *European Economic Review*, 92, 31–46.

Lyons, E. (2017): "Team Production in International Labor Markets: Experimental Evidence from the Field," *American Economic Journal: Applied Economics*, 9, 70–104.

MacLaren, N. G., F. J. Yammarino, S. D. Dionne, H. Sayama, M. D. Mumford, S. Connelly, R. W. Martin, T. J. Mulhearn, E. M. Todd, A. Kulkarni, Y. Cao, and G. A. Ruark (2020): "Testing the Babble Hypothesis: Speaking Time Predicts Leader Emergence in Small Groups," *Leadership Quarterly*, 31, 101409.

Marx, B., V. Pons, and T. Suri (2021): "Diversity and Team Performance in a Kenyan Organization," *Journal of Public Economics*, 197, 104332.

Matsa, D. A. and A. R. Miller (2013): "A Female Style in Corporate Leadership? Evidence from Quotas," *American Economic Journal: Applied Economics*, 5, 136–69.

Miller, M. G. and J. L. Sutherland (2023): "The Effect of Gender on Interruptions at Congressional Hearings," *American Political Science Review*, 117, 103–121.

O*NET OnLine (2022): "Browse by Work Context: Work with Work Group or Team," Retrieved April, 2022, from https://www.onetonline.org/find/descriptor/result/4.C.1.b.1.e?s=1&a=1.

Owan, H. (2014): "How Should Teams Be Formed and Managed?" *IZA World of Labor 2014: 83*.

Pew Research Center (2021): "STEM Jobs See Uneven Progress in Increasing Gender, Racial and Ethnic Diversity," Released April 1, 2021, on https://www.pewresearch.org/science/wp-content/uploads/sites/16/2021/03/PS_2021.04.01_diversity-in-STEM_REPORT.pdf.

Radbruch, J. and A. Schiprowski (2023): "Committee Deliberation and Gender Differences in Influences," Collaborative Research Center Transregio 190, Discussion Paper No. 398.

Remus, R., U. Quasthoff, and G. Heyer (2010): "SentiWS - A Publicly Available German-Language Resource for Sentiment Analysis," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association, Valletta.

Schmid Mast, M. (2002): "Dominance as Expressed and Inferred Through Speaking Time," *Human Communication Research*, 28, 420–450.

Shan, X. (2022): "The Minority Trap: Minority Status Drives Women Out of Male-Dominated Fields," Mimeo.

Spencer Stuart (2021): "2021 U.S. Spencer Stuart Board Index," Released October, 2021, on https://www.spencerstuart.com/-/media/2021/october/ssbi2021/us-spencer-stuart-board-index-2021.pdf.

Terjesen, S., E. B. Couto, and P. M. Francisco (2016): "Does the Presence of Independent and Female Directors Impact Firm Performance? A Multi-Country Study of Board Diversity," *Journal of Management & Governance*, 20, 447–483.

Weidmann, B. and D. J. Deming (2021): "Team Players: How Social Skills Improve Team Performance," *Econometrica*, 89, 2637–2657.

Woolley, A. W., C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone (2010): "Evidence for a Collective Intelligence Factor in the Performance of Human Groups," *Science*, 330, 686–688.