# Pre-Analysis Plan: The Impact of Gender Diversity on Team Communication, Team Performance, and Preferences for Teamwork*

David Hardt† Lea Mayer† Johannes Rincke†

July 21, 2021

## Contents



*The authors would like to thank Ulrich Glogowsky and Hessel Oosterbeek for helpful discussions, and the German Research Foundation (project RI 1959/5-1) and the Hans-Frisch-Stiftung for their financial support.

†University of Erlangen-Nuremberg,
david.hardt@fau.de, lea.mayer@fau.de, johannes.rincke@fau.de

# 1 Introduction

## 1.1 Abstract

We use an online experiment to study how a team's gender composition affects team communication, team output, and preferences for further teamwork. In the first stage of the experiment, randomly composed teams of four meet in an online chat room and work on a series of complex single choice problems under a team piece rate. Team members can communicate via a group audio chat while working on the task. The teams' gender composition varies between all-male, all-female, and mixed (two females and two males). In the second stage of the experiment, each subject meets another subject from a different first-stage team in an online chat room for a short period of time. Subjects then individually state their preference for working on the task individually, or in a team with the other subject. We also elicit the subjects' beliefs about the productivity of individual work and teamwork, and beliefs about team communication. Using digitized data on team communication from the first stage, we ask how a team's gender composition affects team communication, and whether differences in communication translate into differences in team output. Exploiting choice data and beliefs from the second stage, we explore how being assigned to either a gender-diverse or a gender-homogenous team in the first stage affects subjects' beliefs about and preferences for teamwork.

## 1.2 Motivation

Teamwork has become ubiquitous [Bandiera *et al.*, 2013; Owan, 2014] and is widely considered a major component determining the success of organizations [Ichniowski *et al.*, 1997; Ichniowski and Shaw, 1999]. To a large extent, the rise of teamwork reflects the notion that due to complementarities in skills, team workers can be more productive than individual workers alone [Lazear, 1999; Lazear and Shaw, 2007; Mas and Moretti, 2009]. More broadly, following the idea of possible complementarities, the literature has studied performance differences between homogenous and diverse teams along several dimensions other than skills, most prominently culture and language [Hoogendoorn and Van Praag, 2012; Hamilton *et al.*, 2012; Hjort, 2014; Lyons, 2017; Marx *et al.*, 2021] and gender [Adams and Ferreira, 2009; Apesteguia *et al.*, 2012; Hoogendoorn *et al.*, 2013; Adams *et al.*, 2015]. While the literature has made con-

siderable progress in understanding the role of diversity for team performance, important questions remain. First, evidence on the channels through which diversity affects team production and team output is sparse. Second, while many papers have studied the short-term effects of diversity on team performance, we know little about how team diversity affects workers' perception of teamwork as a job amenity [Hamilton *et al.*, 2003; Maestas *et al.*, 2018; Cooper *et al.*, 2021].

We address these gaps by means of an online experiment that induces exogenous variation in the gender composition of teams working on a complex real-effort team task. Our main interest is twofold. First, we study how the gender composition of teams affects team communication, and whether differences in team communication translate into effects on team output. Although the importance of communication for success in teams is widely acknowledged [Hinds and Mortensen, 2005; Woolley *et al.*, 2010], few studies have explicitly considered differences in communication between gender-diverse and gender-homogenous teams. To make progress in this direction, we fully digitize the verbal communication within teams and analyze it qualitatively and quantitatively. Second, we explore how being a member of a gender-diverse team (as opposed to a gender-homogenous team) affects subjects' beliefs about and preferences for teamwork. Our design thus links the literature on how gender diversity affects team productivity to the literature on endogenous team formation [Kuhn and Villeval, 2015; Cooper *et al.*, 2021; Dahl *et al.*, 2021].

In the online experiment, we invite students of a large public university in Germany for a task to be performed online at a specific time. Randomly composed teams of four meet in an online chat room. The randomization ensures that each team consists of two subjects with above-average and two subjects with below-average cognitive skills, where cognitive skills are measured by the student's A-level GPA. The teams' gender composition varies between all-male, all-female, and mixed (two females and two males). The experiment has two stages. In the first stage, teams jointly work on a complex real effort task consisting of ten single-choice problems for 30 minutes. While working on the task, subjects can communicate via a group audio chat. Each team member receives a bonus for each problem provided that all team members mark the correct answer. In the second stage of the experiment, each subject meets another subject from a different first-stage team in an online chat room for a short period of time. Subjects then individually state their preference for working on a similar task as in the first stage individually, or in a team with the other subject. We also elicit the subjects' beliefs about the productivity of individual work and teamwork,

together with beliefs about team communication.

We believe our design has several advantages relative to alternative settings like, for instance, a laboratory experiment or a natural field experiment. First, subjects collaborate with their teammates via an online conference tool. This mode of social interaction has become extremely common in private as well as professional contexts during the COVID-19 pandemic, in particular for our subjects who are university students. The setting of our experiment is thus one that allows subjects to interact in a rather natural way. Second, a big advantage of our design is that we can record the subjects' verbal communication. Relative to previous research, this feature allows us to describe and analyze the role of communication for teamwork in a much more comprehensive way. Importantly, it would have been very difficult to come up with a similarly suitable way of measuring the teams' communication in a natural field experiment.

The project adds to an extensive literature on how the gender composition of teams affects team performance. A first strand of literature [reviewed by Adams *et al.*, 2015] studies the effect of female members in corporate boards on board decision making and firm outcomes. The evidence from this body of literature is mixed, with some papers finding positive effects on firm performance [Campbell and Mínguez-Vera, 2008; Terjesen *et al.*, 2016], while others find none [Chapple and Humphrey, 2014] or even negative effects [Adams and Ferreira, 2009; Ahern and Dittmar, 2012; Matsa and Miller, 2013]. Related to the literature on boards, further work exploiting observational data includes Fenwick and Neal [2001], who use data from a business simulation game and find that team performance is positively related to the number of women in a team. Also analyzing a student business game, Apesteguia *et al.* [2012] find that all-female teams are outperformed by teams of all other gender compositions, including all-male teams.

A second strand of literature uses experimental designs to identify the effect of gender diversity on team performance. Hoogendoorn *et al.* [2013] exploit a field experiment with undergraduate students in business studies who start up a venture as part of their curriculum. Students are randomly assigned to form business teams and run their business for one academic year. Based on 43 business teams, the results indicate that the businesses' sales and profits increase when the share of women increases from a low to an intermediate level. The study also investigates various possible channels through which gender diversity could affect team performance, including complementarities, learning, monitoring, and dealing with internal conflicts. The authors do not find sup-

port for any of these channels. Lamiraud and Vranceanu [2018] also conduct a randomized experiment within a student business game with groups of five and show that all-male and mixed teams perform significantly better than all-female teams. Marx *et al.* [2021] randomize gender diversity in working teams in Kenya and find that gender-homogeneous teams perform significantly better than gender-diverse teams.[1]

While there exists an extensive literature on gender diversity and team performance, the channels through which the gender composition might affect a team's performance have rarely been addressed. Woolley *et al.* [2010] study the collective intelligence of teams and find that an increasing proportion of females and a more equal distribution of conversational turn-taking increases group intelligence. However, these results seem largely mediated by females scoring higher in terms of social sensitivity. Charness *et al.* [2020] focus on the cost of communication. In a part of their laboratory experiment, group members can send unlimited free-form chat messages to assist their group leader in filling out a puzzle. The authors find congestion effects in team communication and document that, as a result, individuals outperform teams. The evidence in Charness *et al.* [2020] is thus in contrast to Hamilton *et al.* [2003] and Cooper and Kagel [2005], who show that teams tend to outperform individuals in solving complex tasks, and suggests that, at least in certain cases, the performance of teams might actually be improved by reducing the amount of communication. Charness *et al.* [2020] also document that groups with a majority of women outperform those with a majority of men. However, the experiment was not designed to study the effect of teams' gender composition, and no evidence is presented on the specific channels through which the effect might work. Related evidence on the link between diversity and communication costs comes from Lyons [2017], who sets up a field experiment on an online market for contract labor to analyze the interaction between teamwork (as opposed to individual work) and national diversity. Studying groups of two workers, the main result of the paper is that teamwork improves performance in nationally homogeneous groups, but decreases performance in nationally diverse groups. The author interprets these findings as evidence for high communication costs in diverse teams counteracting the benefits of teamwork. Keck and Tang [2018] show that groups with female members are less overconfident and therefore more precise in their group confidence judge-

[1] For further references on the link between gender composition and team performance, see Azmat and Petrongolo [2014]. The early psychological literature on the effects of gender diversity in teams is reviewed in, e.g., Bowers *et al.* [2000] and Bell *et al.* [2011].

ments, an effect that appears to be mediated by the group-level communication intensity.

Finally, the project also contributes to an emerging literature on endogenous team formation. Kuhn and Villeval [2015] use a laboratory study following the experimental paradigm of Niederle and Vesterlund [2007] and demonstrate that in a setting without positive synergies from team production, women tend to choose team-based pay more often than men. The gender difference in choices is at least partly driven by women being more optimistic about their prospective teammate's ability.[2] Hamilton *et al.* [2003] study a garment company switching from individual piece rates to team compensation and find that the most skilled workers are disproportionately more likely to join a team voluntarily even though team compensation often harms their earnings. Following up on this striking result, Cooper *et al.* [2021] use a laboratory experiment and show that high ability subjects tend to avoid teams with revenue sharing. When communication is possible, the data suggest that high ability workers join teams because of expected future financial gains from teaching less skilled teammates. Finally, our work also relates to Dahl *et al.* [2021], who demonstrate that male recruits in the military of Norway who lived and worked with women for eight weeks during boot camp have (at least temporarily) more egalitarian attitudes and are more likely to think that mixed-gender teams perform as well or better than same-gender teams.

## 1.3 Research Questions

We investigate the following research questions:

- Does a team's gender composition affect communication among team members?
- Does a team's gender composition affect team performance?
- How does a team's gender composition affect the team members' preference for further teamwork?

[2] Dohmen and Falk [2011] study gender differences in the choice between a fixed wage and several variable pay schemes, including revenue sharing. In our experiment, subjects choose between an individual and a team piece rate.

# 2 Experimental Design

## 2.1 Summary of the Experimental Design

The experiment will be implemented as an online experiment with students of a large public university in Germany.[3] Many students enrolled at the university are registered with a university-wide mailing list used by researchers to invite registered students to surveys and other research-related tasks that can be conducted online. To collect our data, we invite random subsamples of students on the mailing list via email to participate in an online session at a specific time. We instruct subjects in the invitation to make sure that they are in a quiet working space with good online connectivity while participating in their session.

Each session consists of two stages (for experimental instructions, see the Appendix). In stage 1, we form randomly composed teams of four who meet in an online browser-based chat room to work on a complex real effort task for 30 minutes. The teams' gender composition varies between all-male, all-female, and mixed (two males and two females). The real effort task consists of ten single-choice problems. Each team member receives a bonus for each problem provided that all team members mark the correct answer. While working on the task, subjects can communicate via a group audio chat. Stage 1 ends with a short survey that subjects fill out individually. In stage 2 of each session, we form randomly composed pairs of two subjects who have not met in stage 1. Team members meet in an online chat room for a short period of time. During that time, the subjects perform a simple task that requires them to talk to each other. Once the audio chat is closed, we inform subjects about the possibility that they will work on another task similar to the one in stage 1 for 15 minutes, and ask subjects to state their preference for working on the task individually or in a team with the subject they met in the chat room. Before eliciting the preference, we inform subjects about a random draw with three possible outcomes: (a) both subjects who met will work on the task individually, irrespective of their stated preferences; (b) their stated preferences will be implemented; and (c) they will not work on the task at all. We also elicit subjects' beliefs about productivities (individual and team productivity), and beliefs about team communication and interaction should the subject work with the potential teammate.[4] We then im-

[3] Our experiment is programmed with oTree [Chen *et al.*, 2016].

[4] We decided not to elicit preferences and beliefs in the baseline because we were concerned that social desirability concerns and/or experimenter demand would distort the subjects' behavior and communication in stage 1.

plement the random draw regarding the task, and (if determined by the draw) subjects work on the task. The session ends with a brief survey.

## 2.2 Task Stage 1

In stage 1, all teams work on the same set of 10 single choice problems, grouped into two blocks *A* and *B* comprising 5 problems each. We developed the task based on publicly available samples of tasks used to measure applicants' problem solving skills in assessment centers.

At the beginning of each block of problems, subjects obtain access on their screens to complex information material consisting of written text, tables and/or figures. Subjects see the material on their computer screens and get sufficient time to study the material. When the assigned reading time is over, the subjects' screens display the block's first problem, consisting of four written statements on the material presented before. Subjects in a team have three minutes to discuss the material and the statements. We instruct subjects that one out of the four statements is true, and that their task is to mark the correct statement. We also instruct subjects that they receive a bonus for a given problem only if all team members mark the correct statement before the three minutes are up. Subjects can always access the full information material while working on a block of problems, but they cannot return to problems once the time for solving this specific problem is up. During the time windows given for reading the information material and while working on the problems, the audio chat is open. Hence, team members can always communicate while working on the task. Once the three minutes for a problem are up, the next problem (again four statements) is displayed on the screen. Sometimes additional information material is displayed between problems within a block. In those cases, subjects get additional reading time. The timing of the experiment is fixed, and all subjects in a team are redirected to the next page at exactly the same time.

## 2.3 Gender Compositions Stage 1

In stage 1 of each session, subjects are randomly assigned to teams of four. The teams' gender composition varies between all-male, all-female, and mixed (two females and two males).

## 2.4 Gender Compositions Stage 2

In stage 2 of each session, subjects are randomly assigned to pairs of two. The pairs' gender composition varies between all-male, all-female, and mixed (one female and one male).

## 2.5 Elicitation of Preferences for Teamwork in Stage 2

The purpose of stage 2 is to elicit the subjects' preferences for teamwork, conditional on their team's gender composition in stage 1 and their teammate's gender in a possible teamwork in stage 2. To elicit preferences, we proceed as follows. First, we randomly form pairs of two subjects who have not met in stage 1. The pairs meet in an audio chat room for one minute. Each subject's screen displays a 5-digit random number. Subjects are instructed to exchange the numbers and insert the other subject's number in a field. This feature of the design makes sure that the subjects talk to each other and thereby note each other's gender.[5] All subjects progress to the following parts, irrespective of whether they enter the correct number or not. Once the audio chat is closed, we inform subjects about the possibility that they will work on another task similar to the one in stage 1 for 15 minutes, and ask subjects to state their preference for working on the task individually or in a two-person team with the subject they met in the chat room. Before eliciting the preference, we inform subjects about a random draw with three possible outcomes: (a) both subjects who met will work on the task individually, irrespective of their stated preferences; (b) their stated preferences will be implemented as follows: they work as a team if they both indicated this as their preferred option, and they both work individually otherwise; and (c) they will not work on the task at all. We then elicit subjects' preferences for teamwork by asking them whether they would prefer to work on the task individually, or in a team with the subject they met in the chat room. We also elicit subjects' beliefs about their own productivity when working on the task individually, the other subject's productivity when working on the task individually, and team productivity when working with the subject. In addition, we elicit beliefs about team communication and interaction should the subject work with the potential teammate. We then implement the random draw regarding the task.

[5]The communication in this part of the experiment is also recorded. The main reason is that we want to be able to document that the subjects do talk to each other in ways that plausibly enables the subjects to infer the other subject's gender. We do not plan to digitize the communication in this part.

The random draw makes sure that the elicitation of preferences for teamwork is incentive-compatible. It also addresses the concern that subjects might indicate a preference for individual work due to an aversion against a situation that would reveal that one has been rejected as a teammate by the potential partner, or an aversion to reveal to the potential teammate that one has rejected her. Besides explaining the random draw, we explicitly point subjects to the possibility that their stated preferences matter, but that it is also possible that they will work individually no matter what they and the other subject have indicated as their preference, or not work on the task at all.

## 2.6 Experimental Procedures

Subjects are invited for sessions taking part at a specific time. Using a link provided in the invitation email, they enter a webpage. A page with basic instructions informs subjects that they take part in a scientific study, that they will interact with other subjects via an audio chat, and that the audio will be recorded for research purposes. Only subjects who give explicit consent to the recording and to the linking of the data generated during the experiment with administrative data from the university registry can participate. Subjects are then redirected to a microphone test. Next, subjects see a screen with instructions. We inform subjects that they will receive a show-up fee of €10 and that they will work on a series of 10 single-choice problems in a team together with three other subjects they will meet in an audio chat room. We also inform them about a piece rate of €1 for each correctly solved question, and that the piece rate will only be paid if all team members mark the correct statement. We add a statement that further money can be earned in a later stage. We then connect the team members via the chat room for stage 1. Teams have some time to say hello to each other. In the chat room, the subjects of a team are labelled from 1 to 4 (the order is randomly determined). Each team member's number is shown as an avatar in the audio chat and we visualize who is currently speaking. This enables subjects to directly address each other. Next, we instruct the subjects to familiarize themselves with further written technical instructions, including how to reconnect if the internet connection is lost.[6] After that, the information material for the first block of problems is displayed, and teams start working

[6]We inform subjects that all team members' participation in the study will be terminated if one or more team members are not connected to the chat room for more than 90 seconds. In case this happens, all team members receive the show-up fee, but no bonus.

on the problems. With 10 problems and 3 minutes to work on each problem in stage 1, teams work on the task for a total of 30 minutes (net of the time given for reading the information material).

Once the work on the task is completed, the chat room is closed and subjects answer survey 1 individually. The survey contains several items that elicit the subjects' perceptions of team communication and interaction (whether team communication was positive, whether it was cooperative, whether working jointly on task was enjoyable, plus items on quantity of communication, distribution of shares of speech, and interruptions), the belief about how many problems the team has answered correctly, and the belief about how much one has contributed to team output. We also ask subjects to indicate how many of the other subjects in their team they believe were female. In order not to reveal the purpose of the study, we also ask subjects how many of the other subjects in the team they believe are enrolled in certain fields of study, and how many they believe have completed at least two semesters at university.

At the beginning of stage 2, we inform subjects about a further stage of the study, and that they will earn an additional flat payment of €2.[7] We inform subjects that they will meet another subject in the chat room for one minute. Subjects are randomly assigned to form pairs of two. To form pairs, we randomly determine pairs of first-stage teams and randomly match subjects across teams. This makes sure that the newly formed pairs of two consist of subjects who have not met in stage 1. To make sure that subjects in a pair talk to each other, we display a random five-digit number on each subject's screen. Subjects are instructed to fill in the other subject's number in a field on their screen. Once the chat room is closed again, we inform subjects about the possibility that they will work on another task similar to the one in stage 1 for 15 minutes, and ask subjects to state their preference for working on the task individually or in a team with the subject they met in the chat room. The instructions also explain the random draw that determines which of the following scenarios will be implemented: (a) both subjects who met will work on the task individually, irrespective of their stated preferences; (b) their stated preferences will be implemented as follows: they work as a team if they both indicated this as their preferred option, and they both work individually otherwise; or (c) the subjects will not work on the task at all. The subjects then state their preferences for teamwork (binary choice between individual work or teamwork). The subjects' beliefs about their indi-

[7] We inform subjects that they will receive the payment of 2€ only if they complete stage 2 and do not skip any step.

vidual productivity is elicited as follows: we ask subjects to imagine they would work individually on a task similar to the one they have worked on in stage 1, but that the task would comprise 4 blocks of 5 problems each, giving a total of 20 problems. Given a piece rate of € 1 per correctly solved problem, we ask subjects to indicate how many problems they believe they would solve. Similarly, we ask subjects to indicate how many problems (out of 20) they believe the other subject would solve correctly when working on the task individually. Finally, we ask the subjects to state how many problems they believe they would solve when working with the other subject in a team under conditions as in stage 1.[8] In addition, we elicit beliefs about team communication and interaction should the subject work with the potential teammate. For that purpose, we use similar items as in stage 1. Specifically, we use items measuring subjects' beliefs about whether the communications would be positive, whether it would be cooperative, and whether working jointly with the potential teammate on task would be enjoyable. Finally, survey 2 elicits the subjects' perception of the potential teammate's gender. In order not to reveal the purpose of the study, we also ask subjects if they believe the potential teammate is enrolled in certain fields of study and has completed at least two semesters at university.

The random draw regarding the subjects' actual work on the task will be parameterized such that (a) the probability for individual work irrespective of stated preferences is 5%, (b) the probability for the stated preferences to be implemented is 5%, and (c) the probability for no work on the task at all is 90%.

The experiment ends with survey 3, eliciting the Big-5 personality traits. We use a 15-item survey introduced by Gerlitz and Schupp [2005] to measure openness, conscientiousness, extraversion, agreeableness, and neuroticism.

**Timeline**

---

1. Invitation
    - Random subset of individuals on mailing list receives email invitation
    - Invitation contains specific time and date at which the individuals can participate

[8] The elicitation of beliefs uses a frame with 20 problems to enable subjects to express beliefs about relatively small productivity differences.

2. Instructions

    - Subjects receive instructions and are informed about payoff schedule
    - Subjects are randomly assigned into teams (all-female, all-male, and mixed)
    - Audio chat opens

Stage 1

---

3. Real-Effort Team Task

    - Subjects read information material
    - Subjects work on 10 single-choice problems
    - Audio chat closes

4. Survey 1

    - Perceptions regarding team communication and interaction
    - Belief about team output and perception of own contribution to team output
    - Number of females among teammates

Stage 2

---

5. Elicitation of Preference for Teamwork

    - Audio chat opens
    - Subjects meet randomly drawn subject from a different stage 1-team in the chat room ("potential teammate")
    - Audio chat closes
    - Subjects state preference for individual or teamwork with potential teammate (binary choice)
    - Subjects state belief about own productivity if working on the task individually, potential teammate's productivity if working on the task individually, and team productivity if working on the task together with potential teammate

- Subjects state belief about team communication and interaction

6. Survey 2: potential teammate's gender
7. Random draw: (a) individual work, (b) work according to stated preferences, or (c) no work on the task at all
8. Real-Effort Task Stage 2
    - If (a): Individual work
    - If (b): Work according to stated preferences (individual, or teamwork)
    - If (c): No work on the task at all

---

9. Survey 3: Big-5 personality traits
10. Information on total payoff and payment preferences (voucher or bank transfer)

---

## 2.7 Sampling

**Subject Pool** Our sampling frame draws on all subjects on the university's mailing list. To define the pool of subjects to be invited, we first exclude the following groups:

- graduate or exchange students
- students above age 33
- students whose A-level certificate suggests their German might not be fluent (A-level earned abroad and nationality not German)

We aim at a subject pool that allows us to invite random subsamples to experimental sessions such that key individual characteristics, including the A-level GPA, are balanced between female and male subjects. We therefore apply propensity score matching to trim the subject pool by excluding subjects who do not have a sufficiently similar twin of the opposite gender. This removes mostly

**Table 1:** Individual Characteristics of Subjects in Subject Pool

| | Mean | SD |
|---|---|---|
| Female | 0.501 | 0.500 |
| Age | 22.3 | 3.0 |
| Foreign | 0.048 | 0.213 |
| Grade A-level | 232.1 | 62.6 |
| A-level Track=Gymnasium | 0.820 | 0.384 |
| Master Student | 0.194 | 0.395 |
| Arts & Humanities | 0.238 | 0.426 |
| Engineering | 0.211 | 0.408 |
| Natural Sciences | 0.115 | 0.319 |
| Economics & Business | 0.227 | 0.419 |
| Medicine & Law | 0.210 | 0.408 |

This table shows the individual characteristics of the subjects in the subject pool as of June 2021 (=6449)

females with an above-average A-level GPA from the subject pool. After applying these restrictions, the subject pool consists of about 6,400 subjects.

**Characteristics of Subjects in the Subject Pool** Table 1 reports individual characteristics of subjects after applying the aforementioned restrictions. It should be noted that the subject pool changes over time, due to students withdrawing from the mailing list and new registrations.

**Recruitment** For each session, we invite a random subsample of subjects in the pool via email. The email invites the recipient to perform an online task (called an "online minijob") that will last less than one hour. Neither in the invitation email nor during the session, subjects are informed that they take part in an experiment. Subjects are invited for a specific time and receive a reminder email a few hours before the session starts. Subjects who do not respond to the invitation for a given session or show up late may be invited for later sessions.

**Formation of Teams in Stage 1** Our aim is to form teams of four subjects for the first stage of the experiment such that, independent of the teams' gender composition, the teams' composition in terms of subjects' cognitive skills is drawn from the same distribution. Due to possible gender differences in self-selection into participation, it is not sufficient to rely on simple random sampling to achieve this.

To solve the problem, we rely on the GPA of the subjects' university entrance qualification as a comprehensive measure of cognitive skills. To ensure a bal-

anced team-level composition of cognitive skills across all teams, we proceed in two steps. First, for all subjects in the subject pool (after applying the restrictions discussed above), we define an indicator variable for above-average cognitive skills at the level of the individual $i$, $h_i$, and an indicator for females, $f_i$. Based on those indicators, we assign each subject in a given experimental session to one of four bins, $A, B, C, D$. $A$ consists of all subjects with $(h_i = 1, f_i = 1)$. $B$ consists of all subjects with $(h_i = 1, f_i = 0)$, $C$ consists of all subjects with $(h_i = 0, f_i = 1)$, and $D$ consists of all subjects with $(h_i = 0, f_i = 0)$. Second, when randomly assigning subjects to teams, each team is formed by drawing two individuals from high-skills bins $(h_i = 1)$, and two individuals from low-skills bins $(h_i = 0)$. Specifically, when composing an all-female team, we draw from bins $A$ and $C$ (two subjects from each bin), and when forming an all-male team, we draw from bins $B$ and $D$ (two subjects from each bin). When forming mixed teams, we draw either from bins $A$ and $D$ (two subjects from each bin), or from bins $B$ and $C$ (two subjects from each bin), or we draw one subject from each bin. The randomization scheme thus makes sure that all teams are drawn from the same distribution of team-level compositions in terms of cognitive skills, despite the fact that ex-ante, female and male subjects in a given session will generally differ on average in cognitive skills. Subjects who cannot be assigned to a team are informed that they cannot participate in the given session and receive a show-up fee of €2.

**Formation of Teams in Stage 2** In stage 2, subjects are randomly assigned to a potential teammate from another first-stage team. First, we randomly form pairs of first-stage teams. Second, for each pair of teams, we randomly select one team and randomly assign each subject a subject from the other team. If the number of first-stage teams in a session is odd, we randomly select three first-stage teams, then randomly select six subjects from the teams, and randomly assign each of them one of the remaining subjects from a different team. With all remaining first-stage teams, we proceed as described before.

## 2.8 Exclusions

We apply the following rules to define the estimation sample: In stage 1, teams drop out of the experiment if one or more subjects from the team leave the chat room for more than 90 seconds. If teams drop out, we will not consider their data. All subject who finish stage 1 enter stage 2. In stage 2, subjects drop out of the experiment if they do not show up in the chat room, or if they are not

matched to another subject.[9] We let all subjects who take part in the stage-2 matching finish the experiment, irrespective of whether they skip (parts of) the preference and beliefs elicitation. The reason is that we want to collect the Big-5 personality traits in survey 3 from as many subjects as possible. Subjects who skip preference and/or beliefs elicitation questions will not be considered when studying any of the outcomes derived from stage 2. We will also exclude stage-2 observations from subjects who did not enter the correct 5-digit number while being matched to another subjects at the beginning of stage 2.

## 2.9 Planned Sample Size

To test the functionality of the website and to learn about participation rates, we ran several pilot sessions in the spring of 2021. In these pilot sessions, we implemented only stage 1 of the experimental design. While participation rates in our subject pool for tasks that the subjects can do flexibly at any time often reach 60 percent, participation rates in the first pilot sessions were much lower, most of the time well below 10 percent. We suspect this is mostly due to the fact that subjects can participate only at a given time, offering less flexibility. For several reasons, data collection in our experimental design is quite inefficient when sessions are small. First, due to the need to form teams of four of a certain composition in terms of gender and cognitive skills, in small sessions often a large share of the subjects who show up has to be turned down. Second, in stage 2 of the experiment, pairs of subjects are formed across first-stage teams, implying that the inference for estimations using data from stage 2 needs to account for clusters comprising at least two first-round teams. In each session with an odd number of first-stage teams, one of the clusters necessarily comprises three first-stage teams (see below for details on the second-stage sampling). Hence, for a given sample size, in expectation the number of second-stage clusters decreases in the number of sessions needed to collect the data.

To improve the efficiency of data collection, we adjusted the invitation procedure in a stepwise manner (text of invitation email, timing of sessions, timing of reminder, number of subjects invited, etc.). Over several pilot sessions, we managed to increase the participation rate, but we hit a ceiling at about 10 percent. As we do not know how participation rates will evolve when subjects are invited repeatedly, we are unsure how many subjects from the pool we will be

[9] This can happen if, for instance, subjects drop out in survey 1 and do not enter stage 2. We may then be left with an odd number of subjects for the stage-2 matching.

able to recruit for the experiment. We believe a final sample size of between 200 and 400 first-stage teams (or between 800 and 1600 subjects) to be realistic, and report minimum detectable effects for corresponding sample sizes. During the pilot sessions, we only varied the invitation procedure and left the experimental design of stage 1 unchanged. We will use the data collected in those sessions (113 teams in total) when analyzing stage 1 of the experiment. As we did not implement stage 2 in the pilot sessions, the final sample regarding data collected in stage 2 will be smaller than the sample collected from stage 1.

We plan to complete data collection during the year 2021. We plan to recruit new subjects for the subject pool we draw from shortly after the beginning of the fall term 2021 (October 2021). Based on previous recruiting rounds, we expect to be able to recruit about 1,000 new subjects. We may run additional sessions in the fall of 2021 where we invite only newly recruited subjects, following the same sampling procedures as described above.

## 2.10 Statistical Power

In the following, we report estimates of minimum detectable effect sizes for several main outcomes derived from stage 1. Throughout, we assume a significance level of 0.05 and power of 0.8. Estimated means and standard deviations are from pilot sessions. As we do not have any data to estimate means and standard deviations of second-stage outcomes, we do not provide minimum detectable effects for second-stage outcomes. We also do not provide minimum detectable effects for objective first-stage communication outcomes, as we did not digitize the team communication from the pilot sessions yet.

For the **subjective assessment of how positive team communication was in stage 1**, we estimate the team-level mean in gender-homogenous teams to be 4.6 Likert points, with a standard deviation of 0.41. With a sample of $N_{g1} =$ 200 (300, 400) first-stage teams, the minimum detectable effect for differences between teams of a given gender composition is 0.20 (0.16, 0.14).

Regarding the **subjective assessment of how cooperative team communication was in stage 1**, we estimate the team-level mean in gender-homogenous teams to be 4.6 Likert points, with a standard deviation of 0.37. With a sample of $N_{g1} = 200$ (300, 400) first-stage teams, the minimum detectable effect for differences between teams of a given gender composition is 0.18 (0.15, 0.13).

For **how much the subjects liked to work with their teammates in stage 1**, we estimate the team-level mean in gender-homogenous teams to be 3.7 Likert

points, with a standard deviation of 0.56. With a sample of $N_{g1} = 200$ (300, 400) first-stage teams, the minimum detectable effect for differences between teams of a given gender composition is 0.27 (0.22, 0.19).

We estimate the team-level mean of the **number of correctly solved problems in stage 1** in gender-homogenous teams to be 4.0, with a standard deviation of 1.76. With a sample of $N_{g1} = 200$ (300, 400) first-stage teams, the minimum detectable effect for differences between teams of a given gender composition (for instance, mixed vs. all-male teams) is 0.83 (0.68, 0.59).

Regarding estimations at the level of the individual subject, we report minimum detectable effect sizes for gender-specific estimations of the effect of being in a mixed team, relative to being in a gender-homogenous team. Hence, in each estimation, we use about half of the overall sample of subjects.

For the **subjective assessment of how positive team communication was in stage 1**, we estimate the individual-level mean in all-female teams to be 4.6 Likert points, with a standard deviation of 0.65. With a sample of $N_{i1} = 800$ (1200, 1600) subjects overall (about half of them females), the minimum detectable effect for differences between teams of a given gender composition is 0.26 (0.21, 0.18).

Regarding the **subjective assessment of how cooperative team communication was in stage 1**, we estimate the individual-level mean in all-female teams to be 4.6 Likert points, with a standard deviation of 0.61. With a sample of $N_{i1} = 800$ (1200, 1600) first-stage teams, the minimum detectable effect for differences between teams of a given gender composition is 0.24 (0.20, 0.17).

For **how much the subjects liked to work with their teammates in stage 1**, we estimate the individual-level mean in all-female teams to be 3.6 Likert points, with a standard deviation of 0.92. With a sample of $N_{i1} = 800$ (1200, 1600) first-stage teams, the minimum detectable effect for differences between teams of a given gender composition is 0.37 (0.30, 0.26).

Standard deviations in all-male teams are slightly larger, leading to slightly larger minimum detectable effects for male subjects.

# 3 Empirical Analysis

This section describes how we will analyze our data once it is available. All the results will be reported in the paper or appendix. If we run additional specifications, we will label them non-registered or exploratory in the paper.

## 3.1 Balancing Checks and Other Design Checks

### 3.1.1 Balancing Checks

**Team Level** We provide balancing checks at the team level for both stages. Specifically, we test if teams of different gender compositions are balanced according to the following team-level characteristics:

- Average of team members' A-level GPA
- Minimum of team members' A-level GPA (GPA of subject with lowest measure of cognitive skills)
- Maximum of team members' A-level GPA (GPA of subject with highest measure of cognitive skills)
- Average of team members' age
- Minimum of team members' age
- Maximum of team members' age
- Team share of subjects with a university entrance certificate earned at high school ("Abitur")

We tabulate means and standard deviations of the stated variables by gender composition (all-male, mixed, and all-female teams). We also report $p$-values of $t$-tests for differences in means (mixed vs. all-male, and all-female vs. all-male).

**Individual Level** We provide balancing checks at the level of the individual subject for both stages as follows. For the following characteristics

- A-level GPA
- Age
- Indicator for university entrance certificate earned at high school ("Abitur"),

we tabulate means and standard deviations of the stated variables for female subjects by the gender composition of the subject's team (all-female vs. mixed). We also report $p$-values of $t$-tests for differences in means. For male subjects, we report corresponding statistics (all-male vs. mixed teams). We might include

other characteristics to the tables that we do not expect to be balanced at the team level, due to differences between female and male students in the choice of study programs.

### 3.1.2 Attrition Checks

We test for systematic attrition in the following ways:

- If subjects drop out in the first stage, their team chat room is closed, and participation in the experiment is terminated for all subjects in the respective team. We will check if the probability of teams not completing the first-stage task differs between teams of different gender composition. For that purpose, we will consider standardized differences.
- If we find a non-trivial amount of attrition in the second stage (conditional on completing the first stage), we will check whether observable characteristics of individuals who drop out in the second stage are comparable to those individuals who do not drop out. For that purpose, we will consider standardized differences.

### 3.1.3 Check of Subjects' Awareness of Their Team's Gender Composition

We check whether subjects are aware of their team's gender composition in both stages. In the first stage, after the task is completed, subjects are asked individually how many of their teammates they believe were female. From the pilot session, we expect more than 90 percent of subjects to state the correct number. If this pattern is confirmed in the full sample, we will ignore the fact that some subjects are not aware of their team's gender composition. Note that awareness is not necessary for the gender composition to affect our main outcomes in the first stage, as team communication and performance could be driven by a team's gender composition even if the team members are not (fully) aware of it. Awareness of the team's gender composition in the first stage is also not necessary for most of the effects in second-stage outcomes. Only in the estimations that aim at identifying the effect of the first-stage gender composition on the preference for teamwork and beliefs conditional on the potential teammate's gender in the second stage, we rely on subjects being aware of the first-stage gender composition. To the degree that awareness is incomplete, our estimates will identify intention-to-treat effects.

In the second stage, after the preference for teamwork and beliefs have been elicited, subjects are asked individually whether they believe their potential teammate is female. We expect more than 90 percent of subjects to be aware of the potential teammate's gender. If this is confirmed in the data, we will ignore the fact that awareness is incomplete. This will matter only in the estimations that condition on the potential teammate's gender. To the degree that awareness is incomplete, our estimates will identify intention-to-treat effects.

### 3.1.4 Check if First-Stage Task Is Gender Neutral

Before we started to pilot the team experiment, we had a sample of 55 subjects work individually on the first-stage task. The individuals also answered survey questions on the task and on the Big-5 personality traits. The subjects were drawn from the same subject pool as the subjects sampled for the team task. To test if there is a gender difference in individual performance, we regress the number of correctly solved problems on a female dummy. As controls, we include the A-level GPA, an indicator for subjects who have obtained their university entrance certificate at high school (Abitur), age, an indicator for students at Master level, and indicators for the department the student's study program is mainly governed by. To test if there is a gender difference in how the subjects like the task, we use the response to the respective survey question as the dependent variable and re-run the same specification.

With 55 individual observations, the power to detect possible gender effects is quite limited. To assess whether the first-stage task is gender neutral, we additionally consider the difference in team-level performance between all-female and all-male teams (see 3.3 for details on the specification). If the data suggest that teams of the stated gender compositions perform differently on average, we will run further sessions to collect more individual observations.

## 3.2 Text Analysis

Besides subjective assessments of team communication via survey questions, we also aim to describe team communication objectively. For that purpose, we fully digitize the teams' verbal communication.

### 3.2.1 Text Preparation

For each team, the audio recording from stage 1 is transcribed following predefined rules. For each verbal contribution by a subject, the transcript includes the speaker, a time stamp, the spoken text, and any nonverbal expressions (laughing or sighing). The transcript also marks interruptions by other speakers.

The transcripts are read into python for lemmatizing and part-of-speech tagging. Words not recognized by the lemmatizer are extracted and manually categorized into words with spelling errors and words with a lemma that is missing in the list of lemmas. Spelling errors are corrected in the transcript, and missing lemmas are added to the list of lemmas.

Words like "and" or "of" are categorized as stopwords via a corresponding list. For the quantitative text analysis, stopwords are maintained. For parts of the qualitative analysis, in particular the topic analysis (see below for details), stopwords are removed.

### 3.2.2 Quantitative Measures

We use several measures capturing how often certain attributes occur in a team's communication. If indicated, we also use the respective measures at the level of the individual subject.

**Number of words** $M_w$ The number of words gives a simple measure of the extent of information exchanged between team members [Boies *et al.*, 2015; Hansen *et al.*, 2017]. We also consider the number of words spoken at the level of the individual subject.

**Number of contributions** $M_c$ A second measure for the extent of information exchanged is the number of contributions at team level. We also consider the number of contributions at the level of the individual subject.

**Number of questions** $M_q$ The number of questions could point to cooperative communication behavior, but also indicate that the team communication is characterized by uncertainty. For example, Hirschman [1994] finds that more questions are asked in gender-diverse teams, and Mulac *et al.* [2001] argues that questions are a more female language feature.

**Number of Interruptions** $M_i$ Interruptions could indicate that some subjects dominate the team communication. Zimmermann and West [1996] shows differences in interruption behavior of men and women. Blair-Loy *et al.* [2017] also conclude that differences in interruption behavior can exist, at least in certain settings.

**Number Non-Verbal Expressions** $M_l$, $M_s$ The number of laughs is counted by $M_l$, while $M_s$ counts sighs. Those measures are meant to capture the non-spoken sentiment of the team [Del Giudice, 2015].

**Number of Hedges** $M_h$ Hedges are words or phrases that soften a statement ("maybe", "I think"). Mulac *et al.* [2001] argues that hedges are a more female language feature.

In addition to the measures capturing how often certain attributes occur in a team's communication, we also consider measures for how (un)equally the team members contribute to the team's communication. We focus on the number of words spoken by and the number of contributions made by individual subjects and use the team-level Herfindahl index (HHI) to derive the distributional measures. This aspect of our analysis relates our work to Woolley *et al.* [2010], who use the variance in turn-taking and show that teams with a few dominating individuals perform worse.

**Distribution of number of words spoken at team level** $HHI(M_w)$ The variable $HHI(M_w)$ measures how (un)equally distributed the number of words are in a team.

**Distribution of turn-taking at team level** $HHI(M_c)$ The variable $HHI(M_c)$ measures how (un)equally distributed turn-taking is in a team.

### 3.2.3 Sentiment Analysis

Sentiment analysis is conducted to capture the sentiment of a team's verbal communication. For that purpose, we use the German sentiment list SentiWS [Remus *et al.*, 2010]. The list contains 1,644 positive and 1,827 negative base forms and their polarity. In total, the list comprises 31,132 inflections.

Sentiment analysis uses natural language processing. We use a lexical approach combined with selected rules, since there is, to our knowledge, no an-

notated corpus of spoken German. For each word, we check whether it is a sentiment word. If so, we count it as a negative or positive sentiment word and add its polarity to the raw polarity measure.

If negation occurs in a part of a sentence, the polarity of the sentiment words in that part of the sentence are inverted and previously positive counted sentiment words are counted as negative (and vice versa).

Based on pre-tests with annotated corpuses [Clematide *et al.*, 2012; Schmidt and Burghardt, 2018], we use the true forms of words (instead of their lemmas), and the polarity is assigned to the base form or inflection in the SentiWS list.

The sentiment analysis will use the following measures (if indicated, we also use the respective measures at the level of the individual subject):

**Polarity of communication** $M_s$ The polarity measure is derived as the sum of the words' polarities divided by the number of sentiment words. We also consider polarity at the level of the individual subject.

**Ratio of words with positive to words with negative sentiment** $M_{rs}$ The ratio is based on the sentiment of all words on the SentiWS list.

**Variance in sentiment** $M_{vs}$ The polarity of a team member's share of the conversation is derived in the same way as the polarity of the team communication as a whole. From the individual polarity measures, we derive the variance of polarity at team level.

### 3.2.4 Topic Analysis

To study the distribution of topics at team level, we will analyze the communication for each of the 10 problems in the first stage separately.[10] We use Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] to derive the topics. LDA is a generative probabilistic model that represents a document by topic probabilities. Since LDA is an unsupervised learning algorithm, no prior categorization of the data is needed. We plan to assess the performance of the method by propensity score and/or coherence score. If LDA does not lead to satisfying results, we will also consider the Markow Model, a Biterm Topic Model, or another approach. If

[10] We assume that the problems define different topics. In case this distinction turns out not to be useful, we may also consider topics across problems.

none of the models results in meaningful topics, we will follow Gentzkow *et al.* [2019] and extract the topics manually.

As a result of the topic analysis, we will get a number of $K$ topics for each problem. The topics contain a list of words and their probability for the topic. Afterward, we will calculate the probabilities for the communication in each team and problem to belong to the topics. The calculated probabilities are based on the list of words in each topic and the words used by the team members. This will be done at team level as well as for individual subjects. $\psi_{itp}$ is the distribution of topics for speaker $i$ in team $t$ in problem $p$. $\overline{\psi}_{tp}$ is the distribution of the topics for team $t$ in problem $p$.

**Topical concentration index at team level** $M_{th}$ Following Hansen *et al.* [2017], we measure how focused on specific topics a team's communication is. For that purpose, we use a Herfindahl index at team level, which we construct as follows. For each problem and each team, a Herfindahl concentration index over $\overline{\psi}_{tp}$ is calculated. The average from the 10 problem-level indices gives the team's Herfindahl concentration index $M_{th}$.

**Difference over the team members' topic distributions** $M_{tb}$ We also calculate the similarity between the topic distribution of subject $i$ and the team average. For that purpose, we calculate the Bhattacharyya difference for each individual and each problem,

$$\sum_{k=1}^{K} \sqrt{\psi_{itp}^{k} \overline{\psi_{tp}^{k}}}. \tag{1}$$

The average from the 10 problem-specific differences will give the difference of the individual from the team. The average of the individual differences will be used for the team as $M_{tb}$.

Possible alternative difference measures we might consider are the dot product similarity, or the Kullback-Leibler difference.[11]

[11] The dot product similarity is defined as $\sum_{k=1}^{K} \psi_{itp}^{k} \overline{\psi_{tp}^{k}}$, and the Kullback-Leibler difference as $exp\left(-\sum_{k=1}^{K} \overline{\psi_{tp}^{k}} \ln\left(\frac{\overline{\psi_{tp}^{k}}}{\psi_{itp}^{k}}\right)\right)$.

## 3.3 Treatment Effects

### 3.3.1 General Information

**Estimation Strategy** We apply a regression-based estimation approach of the treatment effects. If not indicated otherwise, we use OLS. If we consider multiple outcomes that measure a similar construct, we follow Kling *et al.* [2004] and calculate average (standardized) effect sizes across multiple outcomes. We also report OLS results for each equation.

**Inference** We use Huber-White standard errors. When analyzing outcomes at the level of the individual subjects from stage 1, we will cluster standard errors at the first-stage team level. In stage 2, subjects are randomly assigned to a potential teammate from another first-stage team. For that purpose, we form pairs of first-stage teams and randomly assign subjects to pairs of potential teammates across both first-stage teams. To account for potential correlations originating from first-stage interactions, when analyzing outcomes at the level of the individual subjects from stage 2, we will cluster standard errors at the level of the pair of first-stage teams used to assign subjects into pairs of potential teammates.

In some specifications, we will examine multiple outcomes. In these instances, we correct for multiple hypotheses testing along the lines of the method proposed by List *et al.* [2019].

**Control variables** Gender is an individual characteristic that cannot be varied by the researcher. As a result, the gender composition of teams cannot be varied independently of the team members' remaining individual characteristics. As detailed in 2.7, our sampling frame aims at making sure that the within-team distribution of cognitive skills is identical across teams and independent of the teams' gender composition. However, there are other individual characteristics that correlate with gender, like, for instance, the field of study. To account for the resulting heterogeneity in team characteristics, we include the following control variables in our main team-level specifications for first-stage outcomes:

- Average of team members' A-level GPA
- Minimum of team members' A-level GPA (GPA of subject with lowest measure of cognitive skills)

- Maximum of team members' A-level GPA (GPA of subject with highest measure of cognitive skills)
- Average of team members' age
- Minimum of team members' age
- Maximum of team members' age
- Team share of subjects with a university entrance certificate earned at high school ("Abitur")
- Team share of foreigners
- Team share of Master students
- Vector of variables capturing team shares of students in subject categories

The vector of variables capturing team shares of students in certain subject categories covers rather broad sets of fields of study: Arts & humanities, engineering, natural sciences, economics and business administration, and medicine and law. We expect that including the variables capturing the minimum and maximum of team members' A-level GPA and age will not affect our results. If this is the case, we may also drop these variables from the vector of controls.

In specifications using the individual data, we will use the following control variables:

- A-level GPA
- Age (in years)
- Indicator for university entrance certificate earned at high school ("Abitur")
- Indicator for foreigners
- Indicator for Master student
- Vector of indicators for subject categories

### 3.3.2 Effects on Team Communication

The data from stage 1 of the experiment are used to analyze how a team's gender composition affects team communication. We use two types of outcomes: subjective assessments of team communication elicited via survey questions, and objective measures. We run estimations both at the level of the teams $g$, $g = 1, \ldots, M$, and the level of the individual subject $i$, $i = 1, \ldots, N$.

At team level, we estimate the specification

$$Y_g = \beta_0 + \beta_1 T1_{FM,g} + \beta_2 T1_{FF,g} + X_g'\gamma + u_g, \tag{2}$$

where $Y_g$ is the team outcome of interest, $T1_{FM,g}$ is an indicator for gender-diverse teams in stage 1, and $T1_{FF,g}$ is an indicator for all-female teams. $X_g$ captures the team-level control variables.

Our main communication outcomes for this specification are the quantitative measures $M_w$ (number of words spoken) and $M_c$ (number of contributions). Additionally, we consider as main outcomes the distributional measures $HHI(M_w)$ and $HHI(M_w)$ and the polarity of team communication $M_s$. From the subjective measures, we consider as main outcomes the team averages of the survey responses to the questions how positive the team communication was, how cooperative it was, and whether the subjects liked to work with their teammates (all measured on a 5-point Likert scale). All other communication outcomes are considered secondary outcomes. The topic analysis is considered exploratory in nature entirely, and we may decide not to report the results in the paper.

At subject level, we estimate the specification

$$Y_i = \beta_0 + \beta_1 T1_{FM,i} + X_i'\gamma + u_i, \tag{3}$$

where $T1_{FM,i}$ is an indicator for subjects in gender-diverse teams in stage 1, and $X_i$ captures the individual-level control variables. We run the specification on all subjects, and separately for female and male subjects. Alternatively, we may also run

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

Our main communication outcomes for this specification are the individual-level quantitative measures $M_w$ (number of words spoken) and $M_c$ (number of contributions). From the subjective measures, we consider as main outcomes

the subjects' assessments about how positive the team communication was, how cooperative it was, and whether the subjects liked to work with their teammates. All other individual-level communication outcomes are considered secondary outcomes.

### 3.3.3 Effects on Team Performance

To test if a team's gender composition affects team output, we run specification (2), using the number of correctly solved problems in stage 1 as the outcome.

### 3.3.4 Mediation Analysis

If we find that a team's gender composition affects team output, we will try to shed light on whether communication is a causal mechanism through which the gender composition affects team performance. The mediation analysis is conducted at team level and follows the methods suggested by Imai *et al.* [2011] and Imai and Yamamoto [2013]. The analysis is exploratory in nature entirely, and we may not report any mediation analysis if we do not find effects on team output.

### 3.3.5 Effects on Beliefs about and Preference for Further Teamwork

**Effects on Preference for Teamwork** To analyze how a team's gender composition in stage 1 affects the preference for teamwork in stage 2, we consider the following specifications. First, we analyze how the gender composition in stage 1 affects the preference for working in teams. For that purpose, we use specification (3), where $Y_i$ is an indicator for subjects who have stated a preference for teamwork (as opposed to individual work). To explore gender differences in how the exposure to a gender-diverse team environment in the first stage affects preferences, we also run the regression separately for female and male subjects. Alternatively, we may also run an interacted model following (4).

Second, to explore how the preference for teamwork depends on the potential teammate's gender in stage 2, we estimate, separately for female subjects and male subjects, the interacted model

$$Y_i = \beta_0 + \beta_1 T1_{FM,i} + \beta_2 T2_{FM,i} + \beta_3 T1_{FM,i} \times T2_{FM,i} + X_i'\gamma + u_i, \qquad (5)$$

where $T2_{FM,i}$ is an indicator for individuals who have been paired with a potential teammate of the opposite gender in stage 2. If the coefficients $\beta_1$, $\beta_2$, $\beta_3$

have the same signs in both estimations (i.e., for female and male subjects), we may also, as a secondary specification, estimate the model using all subjects.

**Effects on Beliefs About Productivity** To analyze how a team's gender composition in stage 1 affects beliefs about productivity in stage 2, we consider similar specifications as the ones discussed in the previous paragraph. First, we use (3) and (4) to analyze how the gender composition in stage 1 affects beliefs about output. For that purpose, we use three different variables capturing beliefs as outcomes, $Y_i$: a subject's belief about the number of correctly solved problems if she works individually on the task, her belief about the number of correctly solved problems if the potential teammate works individually on the task, and her belief about the number of correctly solved problems if she works together with the teammate on the task. Second, we will also explore how beliefs depend on the gender composition in stage 1, the potential teammate's gender, and the interaction between the two. For that purpose, we use (5), again separately for female subjects and male subjects, using elicited beliefs as outcomes. If we find for all three outcomes that the coefficients $\beta_1$, $\beta_2$, $\beta_3$ have the same signs in both regressions (i.e., for female and male subjects), we may also, as a secondary specification, estimate the model using all subjects.

**Effects on Communication-Related Beliefs** To analyze how a team's gender composition in stage 1 affects beliefs about team communication in stage 2, we consider the same specifications as discussed in the previous paragraph. The only difference is that $Y_i$ captures a subject's belief about how positive and cooperative the communication would be if the subject would work together with the potential teammate on the task.

**Heterogeneity Analyses** Previous literature has shown that own (perceived) ability relative to the teammates' (perceived) ability predicts subjects' preference for teamwork [Kuhn and Villeval, 2015; Cooper *et al.*, 2021]. We have designed our task such that effective team communication should be key to team output, and although the data from pilot sessions with subjects working individually on the task suggest that cognitive skills (measured by A-level GPA) are positively related to individual output, the effect does not seem excessively high (conditional on further controls, a one-standard deviation improvement in the A-level GPA is associated with an increase in output by about 0.25 standard deviations).

Against this backdrop, we will additionally study the heterogeneity of the treatment effects discussed in the previous paragraphs with respect to cognitive skills, but we do not have strong priors that we will find meaningful heterogeneities. Therefore, we do not commit to report the respective results in the paper.

#### 3.3.6 Additional Regressions

We may use further outcomes for additional exploratory analyses. For instance, we may analyze how a team's gender composition affects the subjects' belief about the first-stage team output and the difference between the belief and actual team output, and the subjects' belief about their own contribution to team output. We may also analyze to what extent gender differences in communication behavior and preferences for teamwork are related to differences in personality traits. To increase statistical power in the analyses of communication and team output from stage 1, we may also run panel estimations with problem fixed effects.

# 4 Variables

## 4.1 Treatment Variables

### 4.1.1 Gender Composition Stage 1

- *Type*: Assigned by design
- *Time of Measurement*: At the beginning of the experiment
- *Measurement*: Composition of first-stage teams in terms of gender
- *Transformation of data to generate variable*: Gender composition at team level is translated into indicators
  - $T1_{FF}$: all-female team
  - $T1_{MM}$: all-male team
  - $T1_{FM}$: gender-mixed team (2 females and 2 males)

### 4.1.2 Gender Composition Stage 2

- *Type*: Assigned by design

- *Time of Measurement*: At the beginning of the stage 2
- *Measurement*: Composition of second-stage pairs in terms of gender
- *Transformation of data to generate variable*: Indicator for female subjects whose potential teammate is male and male subjects whose potential teammate is female, $T2_{FM}$

## 4.2 Outcome Variables

### 4.2.1 Number of Words ($M_w$)

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Words count of transcribed team communication

### 4.2.2 Number of Contributions ($M_c$)

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Count of contributions in transcribed team communication

### 4.2.3 Number of Questions

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Count of questions in transcribed team communication

### 4.2.4 Number of Interruptions

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Count of interruptions in transcribed team communication

### 4.2.5 Number of Laughs

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Count of laughs in transcribed team communication

### 4.2.6 Number of Sighs

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Count of sighs in transcribed team communication

### 4.2.7 Number of Hedges

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Count of hedges in transcribed team communication

### 4.2.8 Distribution of Words at Team Level ($HHI(M_w)$)

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Herfindahl index over $M_w$

### 4.2.9 Distribution of Contributions at Team Level ($HHI(M_c)$)

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Herfindahl index over $M_c$

### 4.2.10 Sentiment

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Overall sentiment score of team communication divided by number of words with sentiment score

### 4.2.11 Relation Sentiment

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Number of the negative sentiment words divided by number of positive sentiment words

#### 4.2.12 Variance Sentiment

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Variance of team members' sentiment score

#### 4.2.13 Herfindahl Index Topics

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Herfindahl index over the teams topic distribution

#### 4.2.14 Bhattacharyya Difference Topics

- *Type*: Team communication measure
- *Time of Measurement*: While teams work on task
- *Measurement*: Transcript of recorded communication
- *Transformation of data to generate variable*: Mean of Bhattacharyya differences over each team members' topic distribution to teams' topic distribution

#### 4.2.15 Perceived Distribution of Speech

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Assessment whether team communication was distributed equally (5-point Likert scale)
- *Transformation of data to generate variable*: Team average of raw data and subject-level raw data

#### 4.2.16 Perceived Sufficiency of Team Communication

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Assessment whether team communication was sufficient to solve problems (5-point Likert scale)
- *Transformation of data to generate variable*: Team average of raw data and subject-level raw data

#### 4.2.17 Perceived Frequency of Interruptions

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Assessment how frequent team members interrupted each other (5-point Likert scale)
- *Transformation of data to generate variable*: Team average of raw data and subject-level raw data

#### 4.2.18 Perceived Sentiment of Team Communication

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Assessment whether team communication was positive (5-point Likert scale)
- *Transformation of data to generate variable*: Team average of raw data and subject-level raw data

#### 4.2.19 Perceived Cooperativeness of Team Communication

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Assessment whether team communication was cooperative (5-point Likert scale)

- *Transformation of data to generate variable*: Team average of raw data and subject-level raw data

#### 4.2.20 Perceived Enjoyability of Teamwork

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Assessment whether subjects liked to work with their teammates (5-point Likert scale)
- *Transformation of data to generate variable*: Team average of raw data and subject-level raw data

#### 4.2.21 Belief Team Output

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Belief about how many problems (out of ten) team has solved correctly (drop-down menu, 0 to 10)
- *Transformation of data to generate variable*: Team average of raw data and subject-level raw data

#### 4.2.22 Belief Own Contribution to Team Output

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Belief about own contribution (in percent) to team output (text field, figures 0 to 100)
- *Transformation of data to generate variable*: Team average of raw data and subject-level raw data

#### 4.2.23 Assessment Gender Composition Team

- *Type*: Survey Item
- *Time of Measurement*: After first-stage task is completed
- *Measurement*: Belief about how many of the teammates were female (drop-down menu, 0 to 3)
- *Transformation of data to generate variable*: Indicator for subjects whose assessment is correct

#### 4.2.24 Team Output

- *Type*: team performance measure, stage 1
- *Time of Measurement*: After task is completed
- *Measurement*: Number of correctly solved problems, 0-10
- *Transformation of data to generate variable*: Raw data

#### 4.2.25 Preference for Teamwork

- *Type*: Survey Item
- *Time of Measurement*: After subjects have met potential teammate in stage 2
- *Measurement*: Stated preference to work alone or with potential teammate (binary choice)
- *Transformation of data to generate variable*: Raw data

#### 4.2.26 Belief Own Second-Stage Output

- *Type*: Survey Item
- *Time of Measurement*: After subjects have met potential teammate in stage 2
- *Measurement*: Belief about own output if subjects would work individually on 20 problems (slider, 0 to 20)
- *Transformation of data to generate variable*: Raw data

#### 4.2.27 Belief Teammate's Second-Stage Output

- *Type*: Survey Item
- *Time of Measurement*: After subjects have met potential teammate in stage 2
- *Measurement*: Belief about potential teammate's output if teammate would work individually on 20 problems (slider, 0 to 20)
- *Transformation of data to generate variable*: Raw data

#### 4.2.28 Belief Second-Stage Output Team

- *Type*: Survey Item
- *Time of Measurement*: After subjects have met potential teammate in stage 2
- *Measurement*: Belief about team output if subject would work with potential teammate on 20 problems (slider, 0 to 20)
- *Transformation of data to generate variable*: Raw data

#### 4.2.29 Belief Second-Stage Cooperativeness

- *Type*: Survey Item
- *Time of Measurement*: After subjects have met potential teammate in stage 2
- *Measurement*: Belief about how cooperative team communication would be if subject would work with potential teammate (5-point Likert scale)
- *Transformation of data to generate variable*: Raw data

#### 4.2.30 Belief Second-Stage Sentiment

- *Type*: Survey Item
- *Time of Measurement*: After subjects have met potential teammate in stage 2

- *Measurement*: Belief about how positive team communication would be if subject would work with potential teammate (5-point Likert scale)
- *Transformation of data to generate variable*: Raw data

#### 4.2.31 Belief Second-Stage Enjoyability of Teamwork

- *Type*: Survey Item
- *Time of Measurement*: After subjects have met potential teammate in stage 2
- *Measurement*: Belief about how much subjects would like to work with potential teammate (5-point Likert scale)
- *Transformation of data to generate variable*: Raw data

#### 4.2.32 Assessment Potential Teammate's Gender

- *Type*: Survey Item
- *Time of Measurement*: After subjects have met potential teammate in stage 2
- *Measurement*: Belief about whether potential teammate is female (binary)
- *Transformation of data to generate variable*: Indicator for subjects whose assessment is correct

### 4.3 Control Variables

#### 4.3.1 Cognitive Ability

- *Type*: Administrative data
- *Time of Measurement*: Before start of experiment
- *Measurement*: A-level GPA in administrative data provided by the university
- *Transformation of data to generate variable*: Team average of raw data, maximum value of raw data at team level, minimum value of raw data at team level, and subject-level raw data. We might re-scale the GPA to adapt to the US standard that higher GPAs indicate better grades.

### 4.3.2 Indicator for German Nationality

- *Type*: Administrative Data
- *Time of Measurement*: Before start of experiment
- *Measurement*: Student's nationality in administrative data provided by the university
- *Transformation of data to generate variable*: Indicator for subjects whose nationality is not German, and team average of the indicator.

### 4.3.3 Age

- *Type*: Administrative Data
- *Time of Measurement*: Before start of experiment
- *Measurement*: Student's age in years as recorded in administrative data provided by the university
- *Transformation of data to generate variable*: Team average of raw data, maximum value of raw data at team level, minimum value of raw data at team level, and subject-level raw data

### 4.3.4 Department Indicators

- *Type*: Administrative Data
- *Time of Measurement*: Before start of experiment
- *Measurement*: Department student's study program is associated with, from administrative data provided by the university
- *Transformation of data to generate variable*: Series of indicators for the university's departments (arts & humanities, engineering, natural sciences, economics and business administration, medicine and law), and team averages of indicators

#### 4.3.5 Indicator for Master Students

- *Type*: Administrative Data
- *Time of Measurement*: Before start of experiment
- *Measurement*: Indicator for students at Master's level, from administrative data provided by the university
- *Transformation of data to generate variable*: Subject-level raw data, and team averages of indicators

#### 4.3.6 Indicator for A-Level Track

- *Type*: Administrative Data
- *Time of Measurement*: Before start of experiment
- *Measurement*: Information on A-level track through which student earned university entrance certificate, from administrative data provided by the university
- *Transformation of data to generate variable*: Indicators for students with the most common A-level track ("Abitur"), and team average of indicator

#### 4.3.7 Big-5 Personality Traits

- *Type*: Survey items
- *Time of Measurement*: Before end of experiment (survey 3)
- *Measurement*: Openness, conscientiousness, extraversion, agreeableness, and neuroticism measured by three items for each dimension following Gerlitz and Schupp [2005]
- *Transformation of data to generate variable*: Aggregation of the three underlying dimensions for each dimension following Gerlitz and Schupp [2005]

# References


ADAMS, R. B., DE HAAN, J., TERJESEN, S. and VAN EES, H. (2015). Board diversity: Moving the field forward. *Corporate Governance: An International Review*, **23** (2), 77–82.

— and FERREIRA, D. (2009). Women in the boardroom and their impact on governance and performance. *Journal of Financial Economics*, **94** (2), 291–309.

AHERN, K. R. and DITTMAR, A. K. (2012). The changing of the boards: The impact on firm valuation of mandated female board representation. *Quarterly Journal of Economics*, **127** (1), 137–197.

APESTEGUIA, J., AZMAT, G. and IRIBERRI, N. (2012). The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science*, **58** (1), 78–93.

AZMAT, G. and PETRONGOLO, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics*, **30**, 32–40.

BANDIERA, O., BARANKAY, I. and RASUL, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, **11** (5), 1079–1114.

BELL, S. T., VILLADO, A. J., LUKASIK, M. A., BELAU, L. and BRIGGS, A. L. (2011). Getting specific about demographic diversity variable and team performance relationships: A meta-analysis. *Journal of Management*, **37** (3), 709–743.

BLAIR-LOY, M., ROGERS, L. E., GLASER, D., WONG, Y., ABRAHAM, D. and COSMAN, P. C. (2017). Gender in engineering departments: Are there gender differences in interruptions of academic job talks? *Social Sciences*, **6** (1), 29.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3** (Jan), 993–1022.

BOIES, K., FISET, J. and GILL, H. (2015). Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity. *The Leadership Quarterly*, **26** (6), 1080–1094.

BOWERS, C. A., PHARMER, J. A. and SALAS, E. (2000). When member homogeneity is needed in work teams: A meta-analysis. *Small Group Research*, **31** (3), 305–327.

CAMPBELL, K. and MÍNGUEZ-VERA, A. (2008). Gender diversity in the boardroom and firm financial performance. *Journal of Business Ethics*, **83** (3), 435–451.

CHAPPLE, L. and HUMPHREY, J. E. (2014). Does board gender diversity have a financial impact? Evidence using stock portfolio performance. *Journal of Business Ethics*, **122** (4), 709–723.

CHARNESS, G., COOPER, D. J. and GROSSMAN, Z. (2020). Silence is golden: team problem solving and communication costs. *Experimental Economics*, **23**, 668–693.

CHEN, D. L., SCHONGER, M. and WICKENS, C. (2016). otree – An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, **9**, 88–97.

CLEMATIDE, S., GINDL, S., KLENNER, M., PETRAKIS, S., REMUS, R., RUPPENHOFER, J., WALTINGER, U. and WIEGAND, M. (2012). MLSA – A multi-layered reference corpus for German sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, University of Zurich.

COOPER, D. J. and KAGEL, J. H. (2005). Are two heads better than one? Team versus individual play in signaling games. *American Economic Review*, **95** (3), 477–509.

—, SARAL, K. and VILLEVAL, M. C. (2021). Why join a team?, forthcoming in: Management Science.

DAHL, G. B., KOTSADAM, A. and ROOTH, D.-O. (2021). Does integration change gender attitudes? The effect of randomly assigning women to traditionally male teams. *Quarterly Journal of Economics*, **136** (2), 987–1030.

DEL GIUDICE, M. (2015). Gender differences in personality and social behavior. In *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, Elsevier, pp. 750–756.

DOHMEN, T. and FALK, A. (2011). Performance pay and multidimensional sorting: productivity, preferences, and gender. *American Economic Review*, **101** (1), 556–590.

FENWICK, D., GRAHAM and NEAL, D. J. (2001). Effect of gender composition on group performance. *Gender, Work and Organization*, **8** (2), 205–225.

GENTZKOW, M., SHAPIRO, J. M. and TADDY, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, **87** (4), 1307–1340.

GERLITZ, J. Y. and SCHUPP, J. (2005). *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP*. DIW Research Notes 2005-4.

HAMILTON, B. H., NICKERSON, J. A. and OWAN, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, **111** (3), 465–497.

—, — and — (2012). Diversity and productivity in production teams. In *Advances in the Economic Analysis of Participatory and Labor-managed Firms*, Emerald Group Publishing Limited, pp. 99–138.

HANSEN, S., MCMAHON, M. and PRAT, A. (2017). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics*, **133** (2), 801–870.

HINDS, P. J. and MORTENSEN, M. (2005). Understanding conflict in geographically distributed teams: The moderating effects of shared identity, shared context, and spontaneous communication. *Organization Science*, **16** (3), 290–307.

HIRSCHMAN, L. (1994). Female–male differences in conversational interaction. *Language in Society*, **23** (3), 427–442.

HJORT, J. (2014). Ethnic divisions and production in firms. *Quarterly Journal of Economics*, **129** (4), 1899–1946.

HOOGENDOORN, S., OOSTERBEEK, H. and VAN PRAAG, M. V. P. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science*, **59** (7), 1514–1528.

— and VAN PRAAG, M. (2012). Ethnic diversity and team performance: A field experiment. In *Tinbergen Institute Discussion Paper 12-068/3*.

ICHNIOWSKI, C. and SHAW, K. (1999). The effects of human resource management systems on economic performance: An international comparison of U.S. and Japanese plants. *Management Science*, **45** (5), 704–721.

—, — and PRENNUSHI, G. (1997). The effects of human resource management practices on productivity: A study of steel finishing lines. *American Economic Review*, **87** (3), 291–313.

IMAI, K., KEELE, L., TINGLEY, D. and YAMAMOTO, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, **105** (4), 765–789.

— and YAMAMOTO, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, **21:**, 141–171.

KECK, S. and TANG, W. (2018). Gender composition and group confidence judgment: The perils of all-male groups. *Management Science*, **64** (12), 5877–5898.

KLING, J. R., LIEBMAN, J. B., KATZ, L. F. and SANBONMATSU, L. (2004). Moving to opportunity and tranquility: Neighborhood effects on adult economic self-sufficiency and health from a randomized housing voucher experiment, working Papers 5, Princeton University, Department of Economics, Industrial Relations Section.

KUHN, P. and VILLEVAL, M. C. (2015). Are women more attracted to co-operation than men? *Economic Journal*, **125** (582), 115–140.

LAMIRAUD, K. and VRANCEANU, R. (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 SHAW, K. L. (2007). Personnel economics: The economist's view of human resources. *Journal of Economic Perspectives*, **21** (4), 91–114.

LAZEAR, E. P. (1999). Globalisation and the market for team-mates. *Economic Journal*, **109** (454), 15–40.

LIST, J. A., SHAIKH, A. M. and XU, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, **22** (4), 773–793.

LYONS, E. (2017). Team production in international labor markets: Experimental evidence from the field. *American Economic Journal: Applied Economics*, **9** (3), 70–104.

MAESTAS, N., MULLEN, K. J., POWELL, D., VON WACHTER, T. and WENGER, J. B. (2018). The value of working conditions in the United States and implications for the structure of wages, NBER Working Paper 25204.

MARX, B., PONS, V. and SURI, T. (2021). Diversity and team performance in a Kenyan organization. *Journal of Public Economics*, **197**, 104332.

MAS, A. and MORETTI, E. (2009). Peers at work. *American Economic Review*, **99** (1), 112–45.

MATSA, D. A. and MILLER, A. R. (2013). A female style in corporate leadership? Evidence from quotas. *American Economic Journal: Applied Economics*, **5** (3), 136–69.

MULAC, A., BRADAC, J. J. and GIBBONS, P. (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research*, **27** (1), 121–152.

NIEDERLE, M. and VESTERLUND, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, **122** (3), 1067–1101.

OWAN, H. (2014). How should teams be formed and managed? IZA World of Labor: 83.

REMUS, R., QUASTHOFF, U. and HEYER, G. (2010). Sentiws-a publicly available German-language resource for sentiment analysis. In *LREC*, Citeseer.

SCHMIDT, T. and BURGHARDT, M. (2018). An evaluation of lexicon-based sentiment analysis techniques for the plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop*, Association for Computational Linguistics.

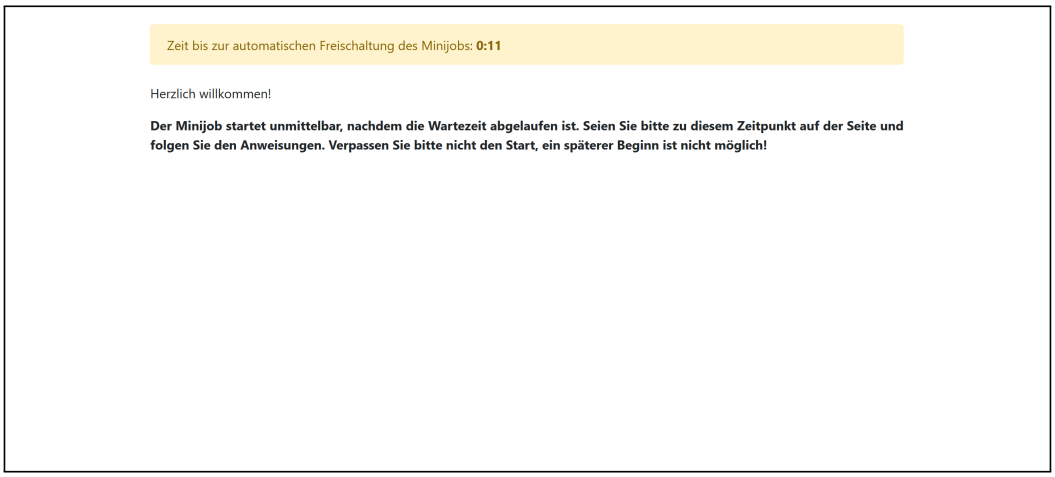
TERJESEN, S., COUTO, E. B. and FRANCISCO, P. M. (2016). Does the presence of independent and female directors impact firm performance? A multi-country study of board diversity. *Journal of Management & Governance*, **20** (3), 447–483.

WOOLLEY, A. W., CHABRIS, C. F., PENTLAND, A., HASHMI, N. and MALONE, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, **330** (6004), 686–688.

ZIMMERMANN, D. H. and WEST, C. (1996). Sex roles, interruptions and silences in conversation. In *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, John Benjamins BV, pp. 211–236.

# Appendices

## A Experimental Instructions




Verbleibende Zeit für Mikrofontest: **1:13**

**Mikrofontest**

Für die Teilnahme an dieser Sitzung benötigen Sie ein Mikrofon.

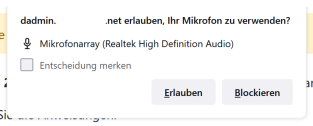
Sie können daher nur teilnehmen, wenn Sie jetzt den Mikrofontest durchführen.

Klicken Sie bitte auf "Mikrofontest starten", um Ihr Mikrofon zu testen.

Mikrofontest starten

---

Verbleibende

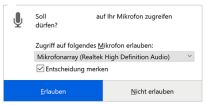



Sie haben noch … ang erfolgreich abzuschließen.

Bitte befolgen Si…


1. **Mikrofonfreigabe**

   Klicken Sie auf **"Erlauben"** oder **"Zulassen"**, wenn Sie eine der folgenden Meldungen erhalten:




oder


2. **Aufhebung der Stummschaltung**

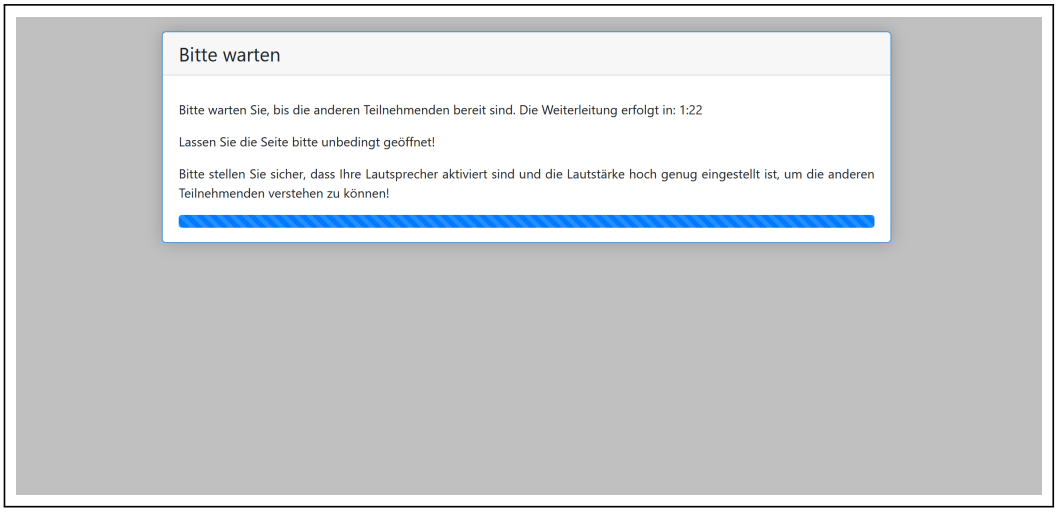
   Ihr Mikrofon ist stumm geschaltet, wenn im Audio-Chatfenster am rechten Rand Ihres Bildschirms folgendes Symbol angezeigt wird:  **Klicken Sie auf das Symbol, um die Stummschaltung aufzuheben.** Die Stummschaltung wurde erfolgreich aufgehoben wenn folgendes Symbol angezeigt wird: 

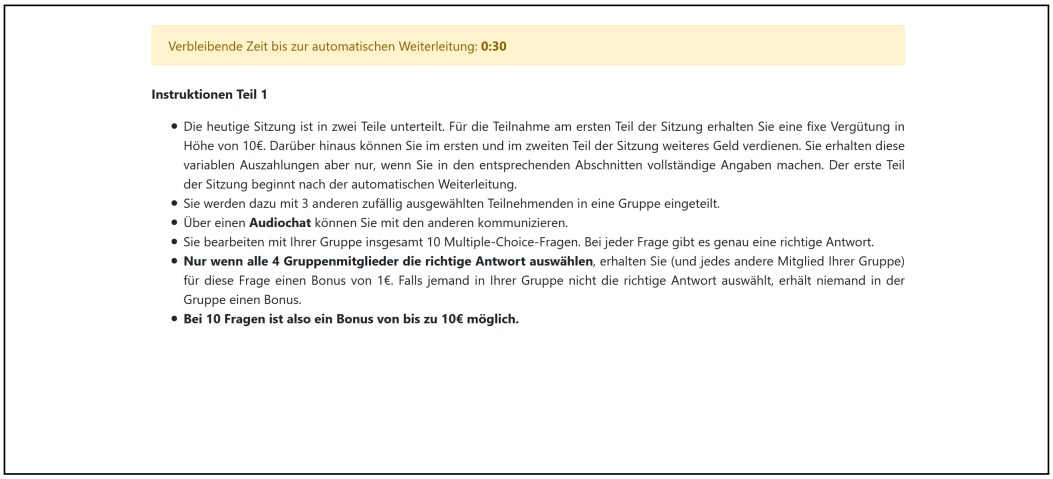
---

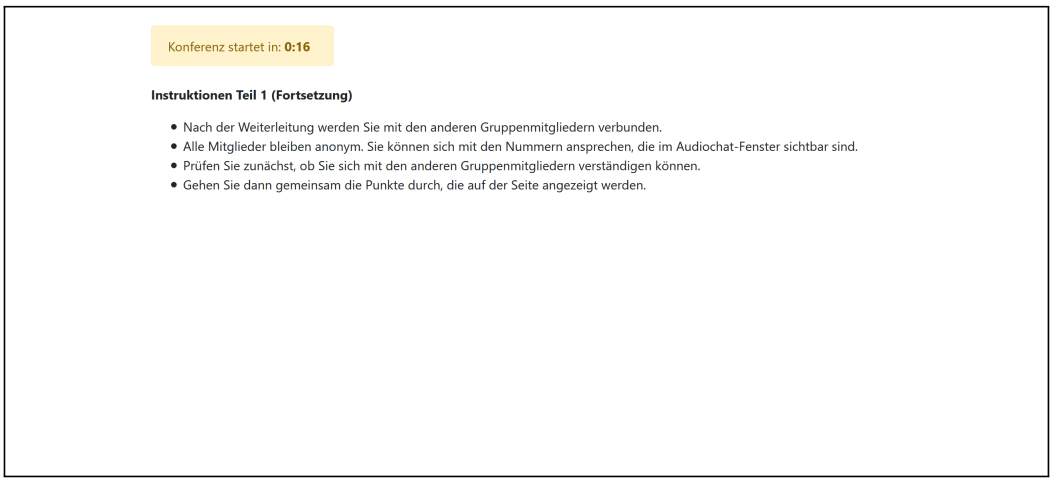
Verbleibende Zeit auf dieser Seite: **0:10**

Der Mikrofontest war erfolgreich.

**Bitte stellen Sie sicher, dass Ihre Lautsprecher aktiviert sind und die Lautstärke hoch genug eingestellt ist, um die anderen Teilnehmenden verstehen zu können!**


Bitte warten
Bitte warten Sie, bis die anderen Teilnehmenden bereit sind. Die Weiterleitung erfolgt in: 1:22
Lassen Sie die Seite bitte unbedingt geöffnet!
Bitte stellen Sie sicher, dass Ihre Lautsprecher aktiviert sind und die Lautstärke hoch genug eingestellt ist, um die anderen Teilnehmenden verstehen zu können!


Verbleibende Zeit bis zur automatischen Weiterleitung: 0:30
Instruktionen Teil 1
Die heutige Sitzung ist in zwei Teile unterteilt. Für die Teilnahme am ersten Teil der Sitzung erhalten Sie eine fixe Vergütung in Höhe von 10€. Darüber hinaus können Sie im ersten und im zweiten Teil der Sitzung weiteres Geld verdienen. Sie erhalten diese variablen Auszahlungen aber nur, wenn Sie in den entsprechenden Abschnitten vollständige Angaben machen. Der erste Teil der Sitzung beginnt nach der automatischen Weiterleitung.
Sie werden dazu mit 3 anderen zufällig ausgewählten Teilnehmenden in eine Gruppe eingeteilt.
Über einen Audiochat können Sie mit den anderen kommunizieren.
Sie bearbeiten mit Ihrer Gruppe insgesamt 10 Multiple-Choice-Fragen. Bei jeder Frage gibt es genau eine richtige Antwort.
Nur wenn alle 4 Gruppenmitglieder die richtige Antwort auswählen, erhalten Sie (und jedes andere Mitglied Ihrer Gruppe) für diese Frage einen Bonus von 1€. Falls jemand in Ihrer Gruppe nicht die richtige Antwort auswählt, erhält niemand in der Gruppe einen Bonus.
Bei 10 Fragen ist also ein Bonus von bis zu 10€ möglich.


Konferenz startet in: 0:16
Instruktionen Teil 1 (Fortsetzung)
Nach der Weiterleitung werden Sie mit den anderen Gruppenmitgliedern verbunden.
Alle Mitglieder bleiben anonym. Sie können sich mit den Nummern ansprechen, die im Audiochat-Fenster sichtbar sind.
Prüfen Sie zunächst, ob Sie sich mit den anderen Gruppenmitgliedern verständigen können.
Gehen Sie dann gemeinsam die Punkte durch, die auf der Seite angezeigt werden.

**Klicken Sie bei technischen Problemen (kein Ton, Audiochat-Fenster nicht sichtbar) hier:** Seite neu laden!

Verbleibende Zeit: **0:11**

Der Audiochat ist nun geöffnet. Im Chatfenster sehen Sie anhand einer Markierung, wer gerade spricht. Bitte machen Sie sich jetzt kurz miteinander bekannt. Die Person mit der Nummer 1 beginnt!

Sie sind: 3


**Klicken Sie bei technischen Problemen (kein Ton, Audiochat-Fenster nicht sichtbar) hier:** Seite neu laden!

Verbleibende Zeit: **0:36**

Sie sind: 3


Sprechen Sie nun gemeinsam die folgenden Punkte durch:

- Wenn Sie den Chat nicht mehr hören oder das Chatfenster nicht mehr sehen können, dann ist Ihre Internetverbindung unterbrochen. **Klicken Sie in diesem Fall bitte auf den roten Button am oberen Rand des Bildschirms.**
- Falls jemand für mehr als 90 Sekunden die Sitzung verlässt, wird die Sitzung für alle geschlossen. Sie erhalten dann alle nur die Teilnahmevergütung in Höhe von 10,00€ (kein Bonus).
- Für jede Frage haben Sie 3 Minuten Zeit. **Um den Bonus für die jeweilige Frage zu erhalten, müssen alle Gruppenmitglieder vor Ende des Countdowns die richtige Antwort auswählen.** Nach Ende des Countdowns können Sie Ihre Antwort nicht mehr ändern.

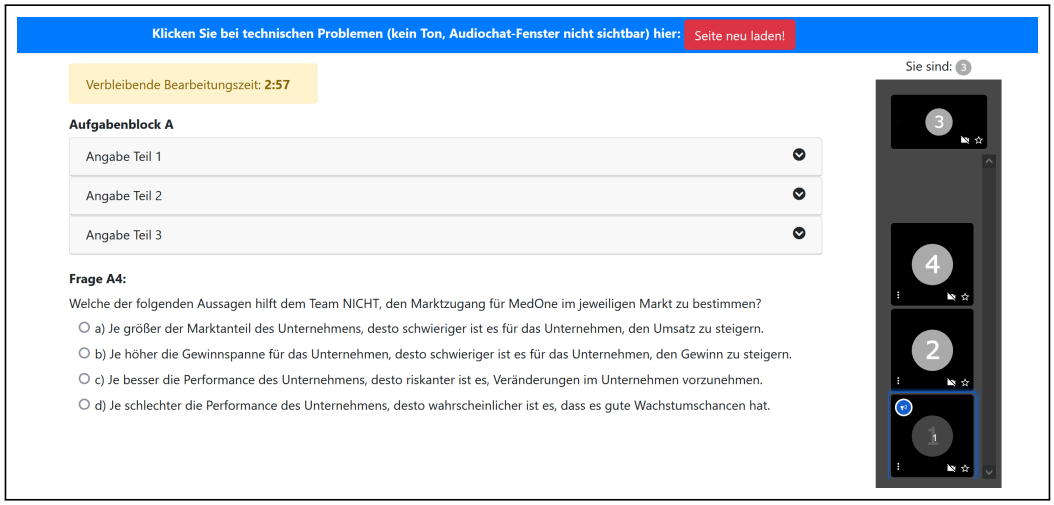
**Klicken Sie bei technischen Problemen (kein Ton, Audiochat-Fenster nicht sichtbar) hier:** Seite neu laden!

Verbleibende Zeit: **0:11**

Sie sind: 3

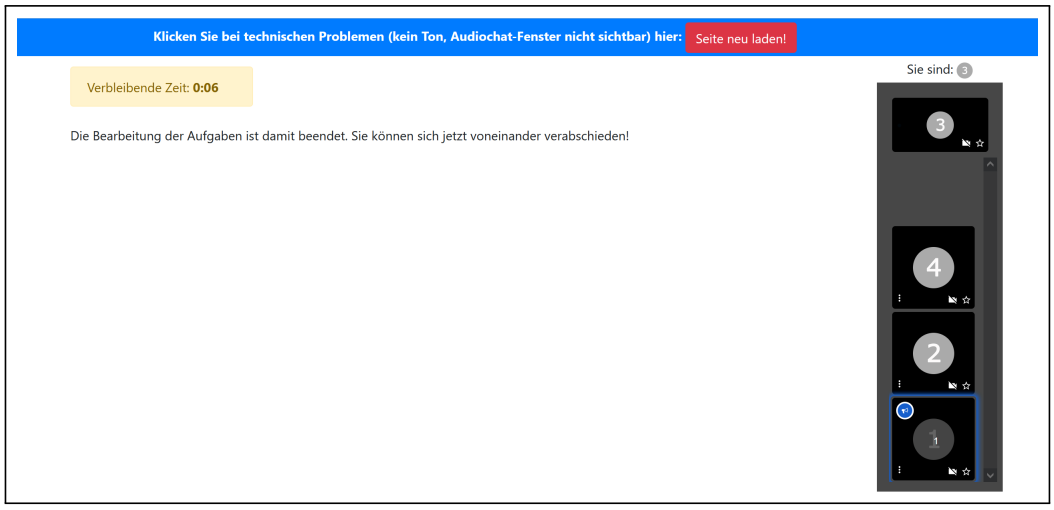

Es beginnt nun die Bearbeitung der Aufgaben. Ihnen wird gleich Material angezeigt, auf das sich die anschließenden Fragen beziehen. Lesen Sie das Material aufmerksam durch. Sie können später jederzeit das Material erneut lesen.

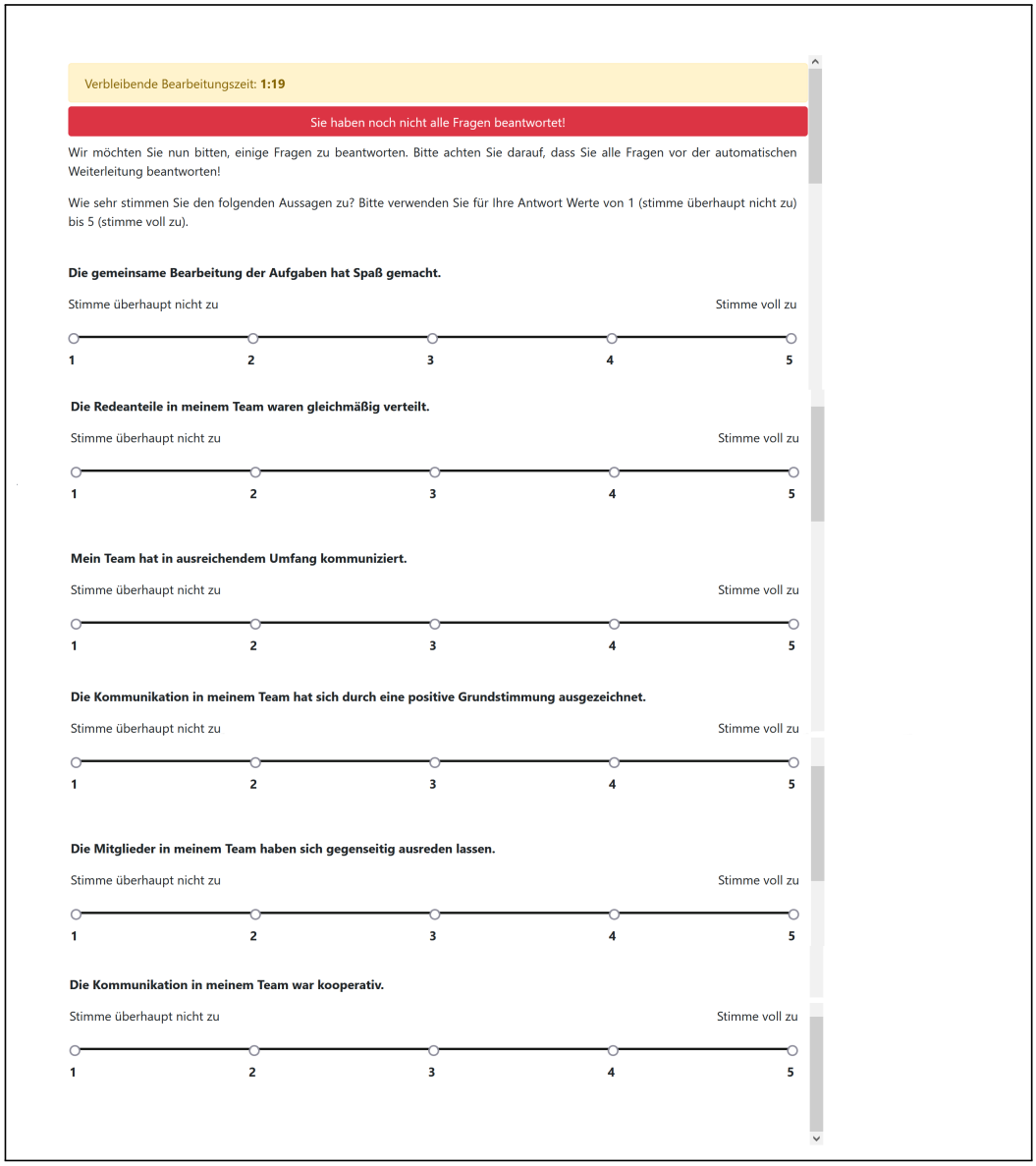
At this point, subject start working on the real effort task (30 minutes plus reading time). While working on the tasks, the subjects can study instructions and information material by opening and closing tabs. In the following, we show a few sample screenshots.

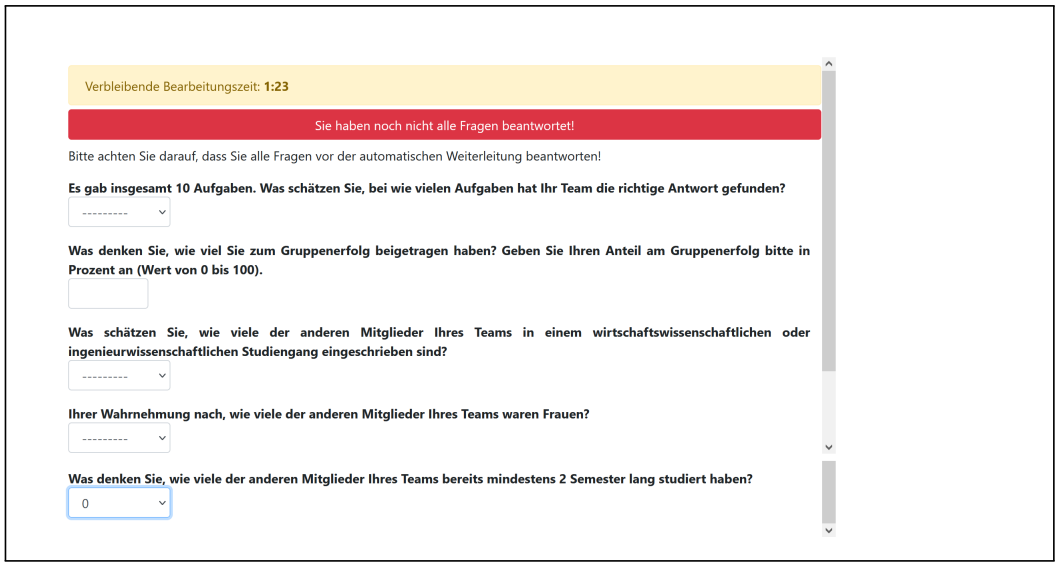


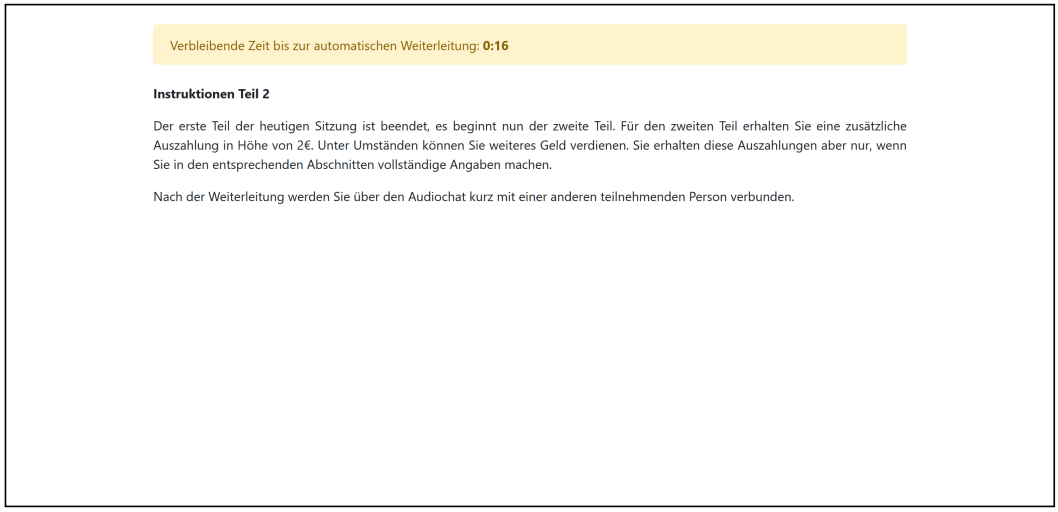


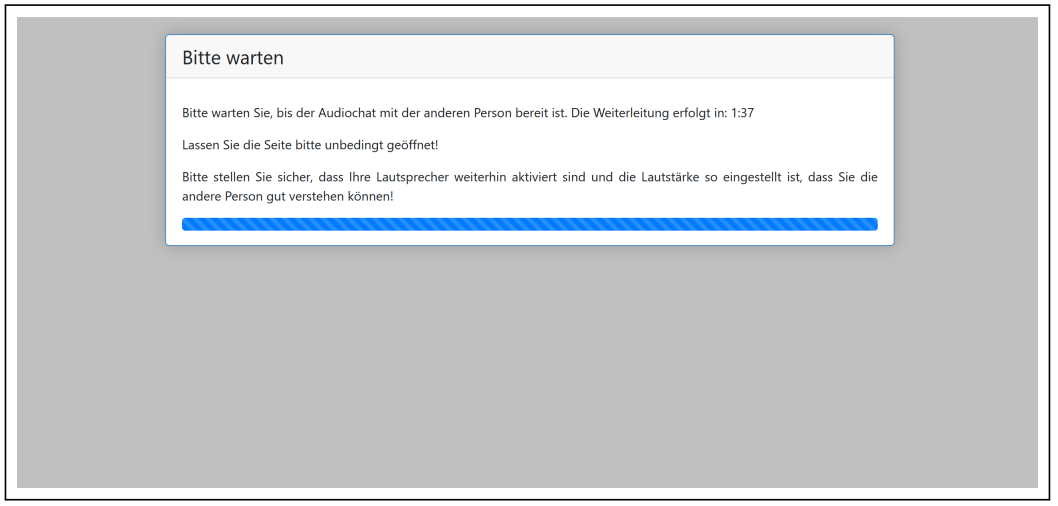
End sample screenshots of task

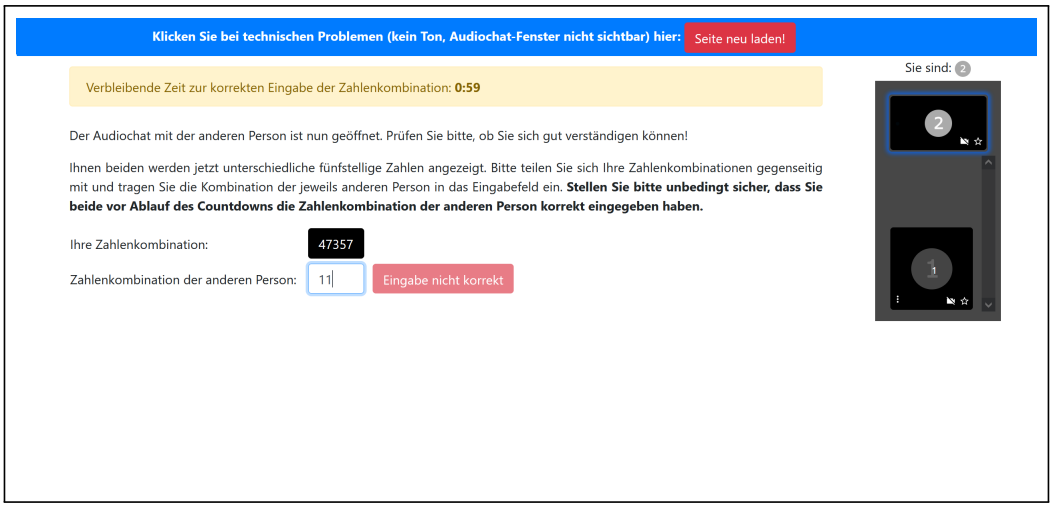

Klicken Sie bei technischen Problemen (kein Ton, Audiochat-Fenster nicht sichtbar) hier:
Seite neu laden!
Verbleibende Zeit: 0:06
Die Bearbeitung der Aufgaben ist damit beendet. Sie können sich jetzt voneinander verabschieden!
Sie sind: 3
3
4
2
1

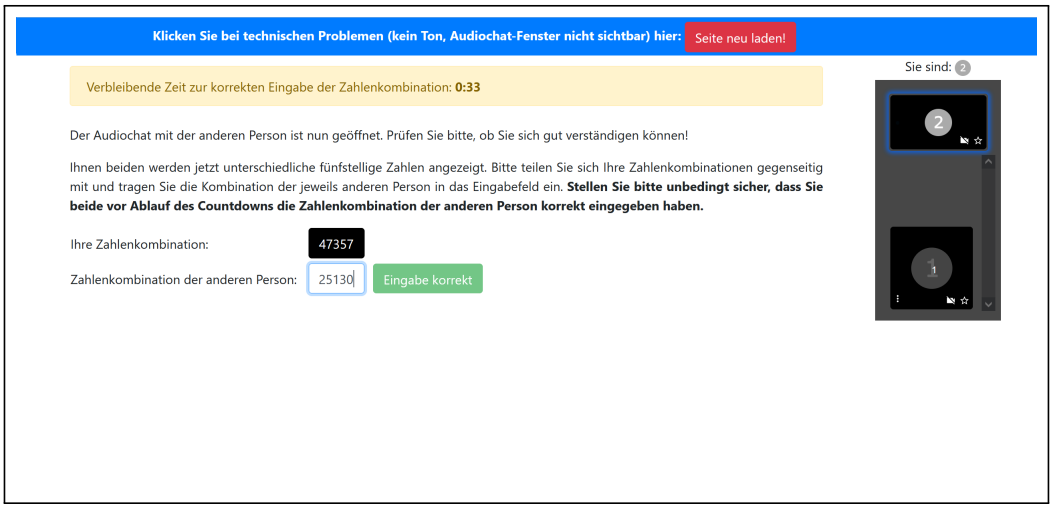

Verbleibende Bearbeitungszeit: 1:19
Sie haben noch nicht alle Fragen beantwortet!
Wir möchten Sie nun bitten, einige Fragen zu beantworten. Bitte achten Sie darauf, dass Sie alle Fragen vor der automatischen Weiterleitung beantworten!
Wie sehr stimmen Sie den folgenden Aussagen zu? Bitte verwenden Sie für Ihre Antwort Werte von 1 (stimme überhaupt nicht zu) bis 5 (stimme voll zu).
Die gemeinsame Bearbeitung der Aufgaben hat Spaß gemacht.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5
Die Redeanteile in meinem Team waren gleichmäßig verteilt.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5
Mein Team hat in ausreichendem Umfang kommuniziert.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5
Die Kommunikation in meinem Team hat sich durch eine positive Grundstimmung ausgezeichnet.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5
Die Mitglieder in meinem Team haben sich gegenseitig ausreden lassen.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5
Die Kommunikation in meinem Team war kooperativ.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5


Verbleibende Bearbeitungszeit: 1:23
Sie haben noch nicht alle Fragen beantwortet!
Bitte achten Sie darauf, dass Sie alle Fragen vor der automatischen Weiterleitung beantworten!
Es gab insgesamt 10 Aufgaben. Was schätzen Sie, bei wie vielen Aufgaben hat Ihr Team die richtige Antwort gefunden?
---
Was denken Sie, wie viel Sie zum Gruppenerfolg beigetragen haben? Geben Sie Ihren Anteil am Gruppenerfolg bitte in Prozent an (Wert von 0 bis 100).
Was schätzen Sie, wie viele der anderen Mitglieder Ihres Teams in einem wirtschaftswissenschaftlichen oder ingenieurwissenschaftlichen Studiengang eingeschrieben sind?
---
Ihrer Wahrnehmung nach, wie viele der anderen Mitglieder Ihres Teams waren Frauen?
---
Was denken Sie, wie viele der anderen Mitglieder Ihres Teams bereits mindestens 2 Semester lang studiert haben?
0


Verbleibende Zeit bis zur automatischen Weiterleitung: 0:16
Instruktionen Teil 2
Der erste Teil der heutigen Sitzung ist beendet, es beginnt nun der zweite Teil. Für den zweiten Teil erhalten Sie eine zusätzliche Auszahlung in Höhe von 2€. Unter Umständen können Sie weiteres Geld verdienen. Sie erhalten diese Auszahlungen aber nur, wenn Sie in den entsprechenden Abschnitten vollständige Angaben machen.
Nach der Weiterleitung werden Sie über den Audiochat kurz mit einer anderen teilnehmenden Person verbunden.


Bitte warten
Bitte warten Sie, bis der Audiochat mit der anderen Person bereit ist. Die Weiterleitung erfolgt in: 1:37
Lassen Sie die Seite bitte unbedingt geöffnet!
Bitte stellen Sie sicher, dass Ihre Lautsprecher weiterhin aktiviert sind und die Lautstärke so eingestellt ist, dass Sie die andere Person gut verstehen können!


Klicken Sie bei technischen Problemen (kein Ton, Audiochat-Fenster nicht sichtbar) hier:
Seite neu laden!
Sie sind:
2
Verbleibende Zeit zur korrekten Eingabe der Zahlenkombination: 0:59
Der Audiochat mit der anderen Person ist nun geöffnet. Prüfen Sie bitte, ob Sie sich gut verständigen können!
Ihnen beiden werden jetzt unterschiedliche fünfstellige Zahlen angezeigt. Bitte teilen Sie sich Ihre Zahlenkombinationen gegenseitig mit und tragen Sie die Kombination der jeweils anderen Person in das Eingabefeld ein. Stellen Sie bitte unbedingt sicher, dass Sie beide vor Ablauf des Countdowns die Zahlenkombination der anderen Person korrekt eingegeben haben.
Ihre Zahlenkombination:
47357
Zahlenkombination der anderen Person:
11
Eingabe nicht korrekt
2
1


Klicken Sie bei technischen Problemen (kein Ton, Audiochat-Fenster nicht sichtbar) hier:
Seite neu laden!
Sie sind:
2
Verbleibende Zeit zur korrekten Eingabe der Zahlenkombination: 0:33
Der Audiochat mit der anderen Person ist nun geöffnet. Prüfen Sie bitte, ob Sie sich gut verständigen können!
Ihnen beiden werden jetzt unterschiedliche fünfstellige Zahlen angezeigt. Bitte teilen Sie sich Ihre Zahlenkombinationen gegenseitig mit und tragen Sie die Kombination der jeweils anderen Person in das Eingabefeld ein. Stellen Sie bitte unbedingt sicher, dass Sie beide vor Ablauf des Countdowns die Zahlenkombination der anderen Person korrekt eingegeben haben.
Ihre Zahlenkombination:
47357
Zahlenkombination der anderen Person:
25130
Eingabe korrekt
2
1

Verbleibende Bearbeitungszeit: **1:41**

Sie haben noch nicht alle Fragen beantwortet!

Im weiteren Verlauf der Sitzung kann es dazu kommen, dass Sie noch einmal für 15 Minuten eine ähnliche Aufgabe wir im ersten Teil der Sitzung bearbeiten. Sie geben nun an, ob Sie lieber allein arbeiten wollen oder im Team mit der anderen Person, die Sie eben im Audiochat getroffen haben.

Über die Bearbeitung der Aufgabe entscheidet ein Zufallsprozess mit drei möglichen Ergebnissen:

Fall A: Sie und die andere Person bearbeiten GAR KEINE zusätzliche Aufgabe.

Fall B: Sie bearbeiten die Aufgabe 15 Minuten lang allein, unabhängig davon, was Sie im Folgenden angeben. Die andere Person arbeitet auch allein.

Fall C: Sie bearbeiten die Aufgabe 15 Minuten lang. Ihre Angabe beeinflusst, ob Sie allein arbeiten oder im Team mit der anderen Person:

- Wenn Sie beide „Team" angeben, dann arbeiten Sie als Team. Sie treffen sich im Audiochat, bearbeiten die Aufgaben zusammen und erhalten einen zusätzlichen Bonus für jede korrekte gemeinsame Antwort.
- Wenn eine/r von Ihnen „Allein" angibt oder beide „Allein" angeben, dann arbeiten Sie beide allein. Sie treffen sich NICHT im Audiochat. Sie erhalten einen individuellen Bonus für jede korrekte Antwort. Wie die andere Person die Aufgabe bearbeitet, spielt für Ihre Auszahlung keine Rolle.

Es ist also möglich, dass Sie beide unabhängig von Ihren Angaben allein arbeiten (Fall B). Falls es nicht zur Teamarbeit kommt, sagt dieses Ergebnis also für die andere Person nichts darüber aus, wie Sie sich entschieden haben. Gleichzeitig gilt auch: Falls es nicht zur Teamarbeit kommt, sagt das nichts darüber aus, wie sich die andere Person entschieden hat. Andererseits kann es sein, dass Ihre Angaben darüber entscheiden, ob Sie allein oder im Team arbeiten (Fall C).

Bitte entscheiden Sie nun, ob Sie lieber im Team oder lieber allein arbeiten würden:

- ○ Team
- ○ Allein

Verbleibende Bearbeitungszeit: **1:05**

Sie haben noch nicht alle Fragen beantwortet!

Bitte achten Sie darauf, dass Sie alle Fragen vor der automatischen Weiterleitung beantworten!

Falls Sie gleich noch einmal eine Aufgabe bearbeiten, wird diese 15 Minuten dauern. Stellen Sie sich jetzt zunächst eine längere Aufgabe vor, die derjenigen ähnelt, die Sie im ersten Teil der Sitzung bearbeitet haben. Die Aufgabe besteht aus 4 Blöcken mit jeweils 5 Teilaufgaben. Insgesamt gibt es also 20 Teilaufgaben. Stellen Sie sich vor, die Bedingungen für die Bearbeitung (Bearbeitungszeit pro Aufgabe, Höhe des Bonus für richtige Antwort usw.) seien wie im ersten Teil der Sitzung.

Bitte achten Sie darauf, dass Sie alle Fragen vor der automatischen Weiterleitung beantworten!

Falls Sie gleich noch einmal eine Aufgabe bearbeiten, wird diese 15 Minuten dauern. Stellen Sie sich jetzt zunächst eine längere Aufgabe vor, die derjenigen ähnelt, die Sie im ersten Teil der Sitzung bearbeitet haben. Die Aufgabe besteht aus 4 Blöcken mit jeweils 5 Teilaufgaben. Insgesamt gibt es also 20 Teilaufgaben. Stellen Sie sich vor, die Bedingungen für die Bearbeitung (Bearbeitungszeit pro Aufgabe, Höhe des Bonus für richtige Antwort usw.) seien wie im ersten Teil der Sitzung.

**Was denken Sie: Wenn Sie die Aufgaben allein bearbeiten würden, wie viele von den 20 Aufgaben würden Sie richtig beantworten?**

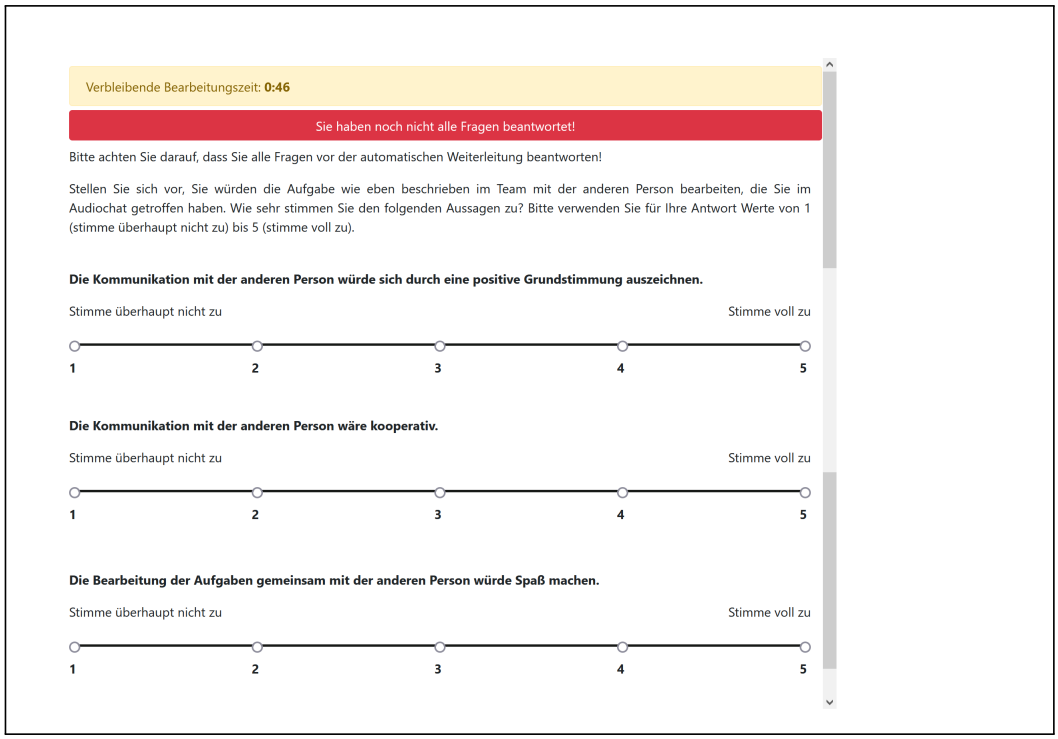
**0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20**

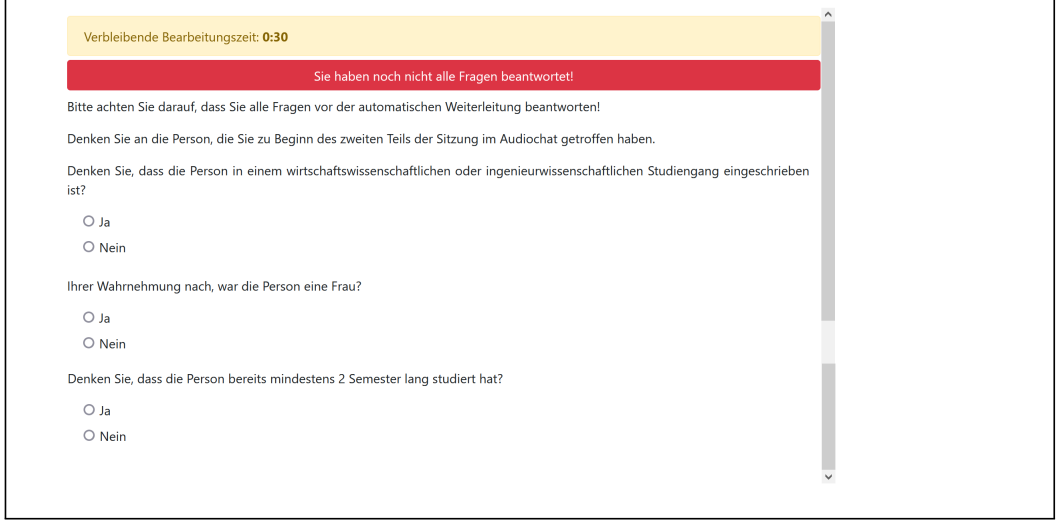
**Was denken Sie: Wenn die Person, die Sie eben im Audiochat getroffen haben, die Aufgaben allein bearbeiten würde, wie viele von den 20 Aufgaben würde sie richtig beantworten?**

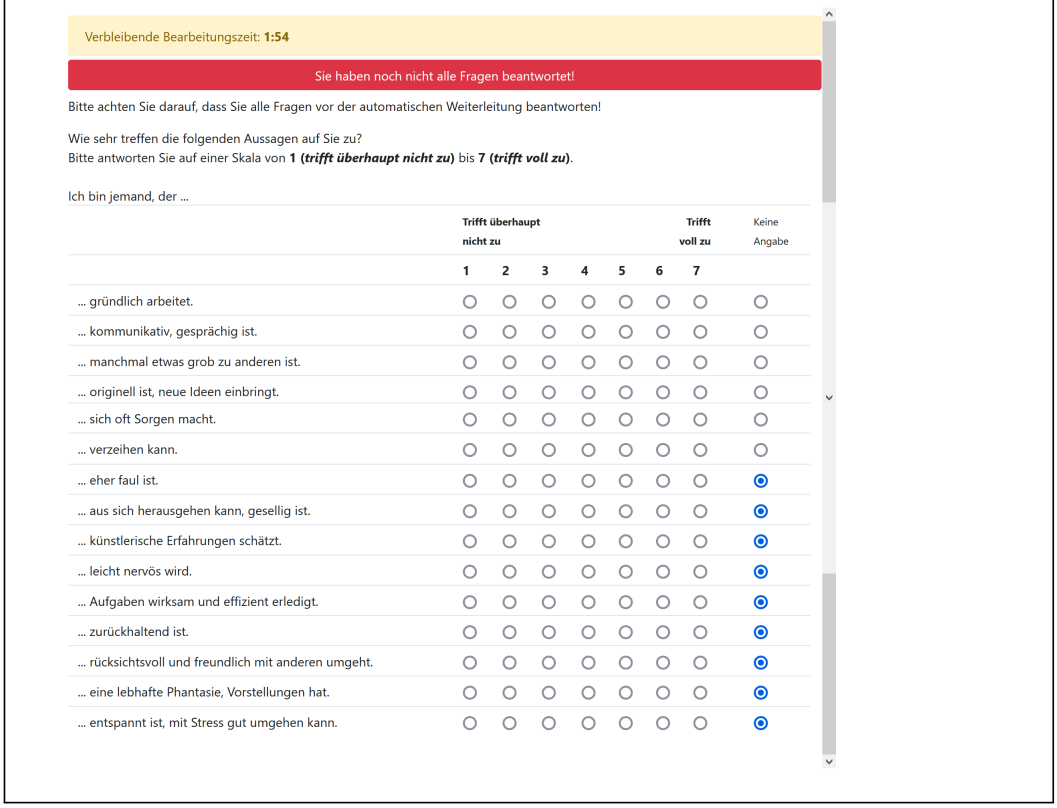
**0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20**

**Was denken Sie: Wenn Sie die Aufgaben im Team mit der Person bearbeiten würden, die Sie eben im Audiochat getroffen haben, wie viele von den 20 Aufgaben würden Sie gemeinsam richtig beantworten?**

**0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20**


Verbleibende Bearbeitungszeit: 0:46
Sie haben noch nicht alle Fragen beantwortet!
Bitte achten Sie darauf, dass Sie alle Fragen vor der automatischen Weiterleitung beantworten!
Stellen Sie sich vor, Sie würden die Aufgabe wie eben beschrieben im Team mit der anderen Person bearbeiten, die Sie im Audiochat getroffen haben. Wie sehr stimmen Sie den folgenden Aussagen zu? Bitte verwenden Sie für Ihre Antwort Werte von 1 (stimme überhaupt nicht zu) bis 5 (stimme voll zu).
Die Kommunikation mit der anderen Person würde sich durch eine positive Grundstimmung auszeichnen.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5
Die Kommunikation mit der anderen Person wäre kooperativ.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5
Die Bearbeitung der Aufgaben gemeinsam mit der anderen Person würde Spaß machen.
Stimme überhaupt nicht zu
Stimme voll zu
1 2 3 4 5


Verbleibende Bearbeitungszeit: 0:30
Sie haben noch nicht alle Fragen beantwortet!
Bitte achten Sie darauf, dass Sie alle Fragen vor der automatischen Weiterleitung beantworten!
Denken Sie an die Person, die Sie zu Beginn des zweiten Teils der Sitzung im Audiochat getroffen haben.
Denken Sie, dass die Person in einem wirtschaftswissenschaftlichen oder ingenieurwissenschaftlichen Studiengang eingeschrieben ist?
Ja
Nein
Ihrer Wahrnehmung nach, war die Person eine Frau?
Ja
Nein
Denken Sie, dass die Person bereits mindestens 2 Semester lang studiert hat?
Ja
Nein



Verbleibende Bearbeitungszeit: **1:54**

Sie haben noch nicht alle Fragen beantwortet!

Bitte achten Sie darauf, dass Sie alle Fragen vor der automatischen Weiterleitung beantworten!

Wie sehr treffen die folgenden Aussagen auf Sie zu?
Bitte antworten Sie auf einer Skala von **1 (*trifft überhaupt nicht zu*)** bis **7 (*trifft voll zu*)**.

Ich bin jemand, der ...

| | **Trifft überhaupt nicht zu** | | | | | | **Trifft voll zu** | Keine Angabe |
|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | |
| ... gründlich arbeitet. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| ... kommunikativ, gesprächig ist. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| ... manchmal etwas grob zu anderen ist. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| ... originell ist, neue Ideen einbringt. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| ... sich oft Sorgen macht. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| ... verzeihen kann. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| ... eher faul ist. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |
| ... aus sich herausgehen kann, gesellig ist. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |
| ... künstlerische Erfahrungen schätzt. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |
| ... leicht nervös wird. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |
| ... Aufgaben wirksam und effizient erledigt. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |
| ... zurückhaltend ist. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |
| ... rücksichtsvoll und freundlich mit anderen umgeht. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |
| ... eine lebhafte Phantasie, Vorstellungen hat. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |
| ... entspannt ist, mit Stress gut umgehen kann. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ◉ |

Verbleibende Zeit bis zur Weiterleitung: **0:06**

Der Zufallsprozess zur möglichen Bearbeitung einer weiteren Aufgabe hat zu dem Ergebnis geführt, dass Sie keine weitere Aufgabe bearbeiten.

Vielen Dank für Ihre Teilnahme!

Bitte entscheiden Sie jetzt, wie Sie Ihre Auszahlung von 10,00€ erhalten wollen. Falls Sie die Option Amazon-Gutschein wählen, werden unsererseits keinerlei Daten an Amazon weitergegeben. Ihren Gutschein erhalten Sie direkt von . Wenn Sie Banküberweisung wählen, müssen Sie auf der nächsten Seite Ihre Kontoverbindung angeben.

- ○ Amazon-Gutschein (Versand per E-Mail innerhalb von 3 Werktagen)
- ○ Banküberweisung (Gutschrift innerhalb von zwei bis vier Wochen)

Weiter