Peer Evaluation and Team Performance: An Experiment on Complex Problem Solving*

John Morgan
University of California, Berkeley
Haas School of Business

Henrik Orzen
University of Mannheim
henrik.orzen@uni-mannheim.de

Dana Sisak†
Erasmus University Rotterdam &
Tinbergen Institute
sisak@ese.eur.nl

This version: January 31, 2024.

Today's employees often work in teams on complex problems. Yet, we know very little about how to incentivize such work. We conduct two laboratory experiments where groups of three work on a complex task and are paid by the quality of their answer. We then study whether a peer evaluation which determines the individual pay share improves group performance. We do so both in an in-person as well as in an online setting. In both settings, overall group performance was not affected by the peer evaluation. Yet, groups behaved differently under peer evaluation: Participants reported higher motivation and groups worked longer and communicated more. We identify two limitations of peer evaluation tied to performance pay. First, behavior may shift towards impressing one's team members, though this does not necessarily improve performance. Second, higher work effort due to the peer evaluation in the presence of time constraints may lead to more timeouts and incompletely worked-out solutions.

JEL Codes: M52, D70, D83
Keywords: Complex Problem Solving, Group Decision Making, Peer Evaluation, Incentives

*We would like to thank Reuben Bauer, Amber Been, Michael van Dijk, David Feitzinger, Bart Francke, Jeroen Frijters, Aniket Godbole, Colin Huliselan, Eline Korndewal, Jeannine van Reeken, Tessa Verloop and Zazie Weiffenbach for excellent research assistance. We also would like to thank Iwan Barankay, Anne Boring, Paul Bose, Suzanne Bijkerk, Philipp Denter, Robert Dur, Florian Englmaier, Chaim Fershtman, Sacha Kapoor, Arjan Non, Simeon Schudy as well as participants of workshops in Amsterdam and Kreuzlingen, seminar participants at Karlsruhe (KIT), Universitat de les Illes Balears, U Warwick, U Köln, Amsterdam (UvA), U Fribourg, U Rotterdam, HU/TU Berlin, U Barcelona, U Innsbruck as well as conference participants at ASSA 2023 New Orleans, SIOE 2023 Frankfurt, U Lüneburg and the CBESS conference in East Anglia. Most of all, we are indebted to Susanne Neckermann, who was heavily involved in the early phases of this project.

†Corresponding author.

1 Introduction

In today's economies, workers are increasingly concentrated in knowledge-based or creative positions. The skills required of these workers differ sharply from those required of prior generations. Mission critical tasks often involve working in teams to solve unique and complex problems. These problems range widely from new product development to the improvement of processes and forecasting techniques. A key characteristic of these types of problems is that individuals need to think creatively and out of the box. Typically, no clear solution procedure exists.

An important question is how to incentivize such teams. Firms differ in their use of incentives in practice. Increasingly, firms use group incentives such as profit sharing (see, e.g., Zenger and Marshall, 2000), but team incentives may fail to motivate because of free-rider problems. Attempts to compensate a worker for her individual contribution to the team require detailed knowledge of the team process, which is often hard for an employer to acquire directly. In practice, many firms use peer evaluations where team members rate each other's contribution to the outcome of the group (e.g., Boyle, 2013, Bracken and Rose, 2011). Edwards and Ewen (1996) report that 90% of Fortune 1000 firms and 18% of American Management Association (AMA) member companies that responded to the survey had implemented some version of peer evaluation. While the information gained through peer evaluations may sometimes be drawn on for developmental and confidential feedback only, many firms utilize this information for performance management. In fact, Edwards and Ewen (1996) report that more than 90% of those AMA member companies that run peer evaluations also use them to determine pay. Whether tying peer evaluations to pay is a good idea is a much-discussed question in the human resource literature (Edwards and Ewen, 1996, Bohl, 1996, Coates, 1998, DeNisi and Kluger, 2000).[1]

While the experimental literature has demonstrated the effectiveness of incentives in increasing performance in simple, individual tasks, complex tasks in group settings have received less attention.[2] We are aware of only two experimental studies that look at the effect of group incentives in

[1] One important distortion discussed is the so-called likability bias, as in for example Sonnentag (1998) and Love (1981), where a more likable colleague receives a more favorable evaluation. Another worry is that workers distort their evaluation of their peers to game the system to their own advantage (Edwards and Ewen, 1996).

[2] See e.g. Bandiera et al. (2021) for a meta study on the effect of incentive pay on individual performance. Incentives may not always be positive. The literature has documented that they can reduce performance by crowding-out intrinsic

complex tasks, Ramm et al. (2013) and Englmaier et al. (2017). Neither examines the effects of peer evaluations. Our paper uses controlled laboratory experiments to study whether adding a peer evaluation to group performance pay increases group performance in a complex task.[3]

To investigate this question, we performed a series of laboratory experiments where groups of paid volunteers had to solve complex problems under group incentives only as well as group incentives augmented with a peer evaluation to determine the individual pay share. To capture problems relevant in business contexts, we chose "guesstimation" tasks, also known as Fermi problems, as our representation of complex problems. Guesstimation tasks pose the challenge of estimating some unknown, and often unusual, quantity.[4] Success in the task derives from chaining together a string of known or easily estimated quantities in a logical, but non-routine, way to estimate the unknown quantity. These types of questions resemble tasks, for example, in management and consulting jobs, such as estimating the potential demand and thus profitability of a new product (Anderson and Sherman, 2010). Consequently many firms use these tasks in interviews of employees, seeing them as measuring critical problem solving abilities needed for their workforce.[5] Guesstimations are also used in higher education. Solving guesstimations promotes the learning of several "21st century skills" such as developing strategies to solve complex problems and promoting collaboration and communication skills (Ärlebäck and Albarracín, 2019). Importantly for our purposes, guesstimations can be chosen to have definite, known, answers, so it is possible to grade performance in an objective fashion.

motivation (e.g. Deci et al., 1999, Frey and Jegen, 2001), induce sabotage by peers (e.g. Harbring and Irlenbusch, 2011, Carpenter et al., 2010) or lead to choking under pressure (e.g. Ariely et al., 2009). There is also a small recent literature examining the effect of incentives on individual creativity, arguably a complex task, with mixed results (see e.g. Bradler et al., 2019, Eckartz et al., 2012, Laske and Schröder, 2017 or Charness and Grieco, 2019). Furthermore, group incentives in simple tasks and the free-rider problem have been studied for example in Fehr and Gächter (2000) in a public goods setting and Nalbantian and Schotter (1997) as well as van Dijk et al. (2001) in an organizational setting. Delfgaauw et al. (2020), for example, study team incentives in a field experiment.

[3]Our research also relates to the small empirical literature studying peer evaluations or peer reviews in organizational research. Balietti et al. (2016), Teplitskiy et al. (2019), Bagues et al. (2017) and Huang et al. (2019) study determinants of peer reviews. Gall et al. (2023) study the effectiveness of subjective performance evaluation on leadership effectiveness in a field setting. Most closely related, Carpenter et al. (2010) study, using a laboratory experiment, the detrimental effects of a peer evaluation in a tournament setting, where competitors are asked to evaluate each other's work and this evaluation determines relative pay.

[4]For example, one of the guesstimation tasks we use in our experiment asks how many litres of toothpaste are consumed in the United Kingdom in one year.

[5]See also https://www.wired.com/2014/08/how-to-solve-crazy-open-ended-google-interview-questions/ for the case of Google. https://www.consultingcase101.com/tag/estimate-or-guesstimate/ shows a database of guesstimation questions that have been used in job interviews.

The flow of our experiment was roughly as follows: We first asked each participant to solve a guesstimation problem and use this data as an ability measure. Next, we grouped her with two other participants, under a given incentive structure, and presented the team with a new problem. Within this group phase, the individual team members were first given a time budget to do the task on their own and then an additional time budget for a group discussion, which resulted in the final team answer. Thus, each group produced three individual answers (as such unincentivized) and one final group answer (incentivized depending on treatment). There was no structure on group discussion. We provided participants with a sheet on which they could write down the steps used in deriving a given guesstimate. We then rematched participants to new teams and repeated this group exercise with a new question. Throughout each session, we kept the incentive scheme constant, and we therefore study how performance varies with incentives between subjects.

We conducted this experiment in two different settings. Our first study (Study 1) considers a classical setting of face-to-face communication and interaction. Teams were seated around a large table, communicated with each other freely, and worked on the guesstimation task using pen and paper. Our follow-up study (Study 2) tests the robustness of our findings to a setting in which teams interact anonymously through computer terminals using a chat application. Our results are remarkably robust across these two settings.

Our main findings are:

1. In both studies, group guesses outperformed the average individual in the preparation phase. Groups were thus valuable for a principal who cannot judge guess quality. At the same time, groups performed worse than the *best* team member. Groups were thus not able to recognize (and improve on) the best individual guess, at least on average.

2. Adding a peer evaluation to group incentives increased self-reported motivation to impress ones' peers and turned the perceived group atmosphere more competitive in both studies. It also increased work time and groups communicated more. At the same time, it led to more timeouts and missing guesses. Ultimately, however, overall group performance was not significantly affected by the addition of a peer evaluation in both studies. Final group guesses were as accurate under the peer evaluation as under group incentives only.

3. Consistent with our first finding, individual guess quality in the preparation phase did not significantly predict success in the peer evaluation. Instead, more easily observable measures of group contribution such as the uniqueness and number of steps on the preparation phase answer sheet or how much a team member contributed to group communication were significantly related to winning the peer evaluation.

Our findings contribute to our understanding of whether a peer evaluation should optimally be tied to performance pay, a question debated in the literature. While likability and gaming issues, previously identified as impairing the effectiveness of peer evaluations (Edwards and Ewen, 1996, Coates, 1998) were minimized in our experimental setting, we identify two further limitations. First, when peers are not fully able to assess individual performance due to imperfect performance feedback, behavior may shift towards impressing one's team members through more extensive preparation efforts and increased communication efforts that are then rewarded by peers, though they do not necessarily improve performance. Second, a higher work effort and motivation due to the peer evaluation in the presence of time constraints may result in more timeouts and incompletely worked-out solutions. In our experiment, participants worked harder and longer under peer evaluation, but at no additional benefit to the employer in terms of increased group performance.

In light of persisting gender gaps on the labor market, it is important to understand gender differences in the response to incentives. Bandiera et al. (2021) show in a meta-study that males and females do not differ in their response to individual performance pay. Much less is known about complex group tasks and incentives relying on subjective performance evaluation by peers. Communication[6], and especially leadership[7] and self-promotion behavior[8] in a group, will likely determine individual success. A peer evaluation may also induce a more competitive environment.[9]

[6]Hardt et al. (2023) document gender differences in communication behavior and gender composition effects in groups working on a team task.

[7]Chakraborty and Serra (2023) and Born et al. (2022) document gender difference and gender composition effects in willingness to lead.

[8]Isaksson (2018) shows that females are more reluctant to claim credit for their work. Coffman et al. (2021) show that male and female contributions to teams are not equally rewarded in gender-stereotyped task. Exley and Kessler (2022) and Lerchenmueller et al. (2019) document gender differences in self-promotion.

[9]Niederle and Vesterlund (2011) survey the literature on gender differences in competitiveness which tend to result from differences in overconfidence and attitudes toward competition. Ariely et al. (2009) show that high-stakes incentive pay may cause choking under pressure. While they do not find gender differences in choking, later studies do (Bracha and Fershtman (2013), Cahlíková et al. (2019), Cohen-Zada et al. (2017), Azmat et al. (2016), and Cai et al. (2019)).

Gender differences in behavior, beliefs, and preferences may imply that men and women react differently to the introduction of a peer evaluation.

In an ex-post analysis, we identified a strong gender effect in the response to peer evaluations in Study 1. Females facing peer evaluation in addition to group incentives finished the preparation phase with more missing guesses and guesses of lower quality than females facing group incentives only. Females under peer evaluation thus entered the group phase with a disadvantage and ultimately female-majority teams underperformed relative to female-majority teams under group incentives only. At the same time, we did not find any evidence of discrimination by gender in the peer evaluation nor of a differential treatment effect during the group phase. We found no significant effect of peer evaluations for males and male-majority teams.

We designed Study 2 also with the aim of shedding further light on the robustness of this ex-post finding. By making the interaction anonymous in Study 2, and thus gender unobservable, we rule out any effect working through discrimination based on gender in the group phase or gender salience. Instead, we focus on channels that work through differences in beliefs/confidence or preferences, as for example an aversion to competition which may cause choking under pressure. We also study whether the gender stereotype of the task matters by choosing a gender neutral guesstimation and one that invokes a male stereotype. We do not replicate the gender effect of Study 1. The differential effect of peer evaluations on men and women, if relevant, seems to be context dependent and requires further study.

Next, we describe the experimental design and the rationale for key design decisions. In Section 3 we present the results of both studies. Section 4 concludes.

2 Experimental design

We present results from two complementary studies. Study 1 took place in-person—teams worked on the guesstimation problem face-to-face. Study 2 took place at computer terminals—teams worked anonymously via chat. Study 1 took place at Erasmus University Rotterdam and consisted of 11 sessions each lasting about 120 minutes. We obtained a total of 171 participants from the econlab participant pool of the Erasmus School of Economics. Study 2 consisted of 20 additional sessions

conducted at the University of Mannheim (mLab) and the University of Heidelberg (AWI Lab). These sessions lasted about 90 minutes each, and a total of 201 participants were recruited. AWI Lab volunteers were recruited via SONA, while for econlab and mLab the ORSEE recruiting software was used (Greiner, 2015). Nobody participated more than once.

We began each session with instructions that contained a worked-out example of a guesstimation question.[10] In Study 1, we provided participants with a fact sheet that included population numbers to level the playing field. In Study 2, participants were allowed to use the internet, so this was not necessary. To aid their calculations, we provided them with calculators. In Study 1, participants were informed that they would be videotaped. In the supplementary appendix we provide both sets of instructions.

Each experimental session consisted of three parts. In Part 1, participants worked on a guesstimation question individually. In Part 2, we formed teams, and these teams solved guesstimation questions under a treatment-specific incentive scheme. In Part 3, participants filled out a final questionnaire. In each session, we kept the incentive scheme constant throughout Part 2 across different rounds of guesstimation questions. We thus identify treatment differences in a between-subjects design.[11] We now present more details on these three parts.

The Part-1 task differed by study, but was the same for all participants within a study. We used performance in this task as a proxy for individual ability at solving guesstimations. A participant's earnings in this part depended on the percentage error of their guess (the percentage deviation from the correct value). We divide these percentage errors into six bands, each with a different payoff. If the answer was within 10% of the correct answer, the full prize of 10 euros is awarded. If the guess was off by plus/minus 20%, 40%, 60%, or 80%, the guesser received 80%, 60%, 40% or 20% of the prize. For example, an individual that submitted a guess that was within 40% of the true answer received $0.6 * 10 = 6$ euros. Participants received no feedback about the quality of their work at any point during the experiment. In Study 2, participants could also earn a small extra amount of

[10] In Study 1, we also taught participants a rule of thumb—given confidence bounds of an item to be estimated, the geometric mean of the bounds often proves to be a good estimate. Both the worked-out example for this study and the rule of thumb were taken from the book "Guesstimation: Solving the World's Problems on the Back of a Cocktail Napkin" by Weinstein and Adam (2008).

[11] We chose a between-subjects design to eliminate contamination in behavior between treatments through learning or confusion.

money by submitting their guess earlier. We implemented this to create salient opportunity costs of time.

Following the individual guesstimation, in Part 2, participants were randomly assigned into groups of three and had to solve a guesstimation problem as a team. In Study 1, each team was sent to a separate room to work on a guesstimation question. Teams had 15 minutes at their disposal to come up with a final answer. We increased this time budget to 22 minutes for Study 2 to account for more cumbersome communication by chat as well as the opportunity to use online search in answering the guesstimation questions.

Part 2 was split into two phases. We will refer to the first of these as the *preparation phase*. Team members were given time to work on the guesstimation question individually. We encouraged them to write down an answer and indicate the steps by which they arrived at their answer by filling out a worksheet[12]. They were made aware that the completion of this worksheet was not in itself incentivized. We added this phase as team members in firms usually have some prior knowledge relevant to the task from, for example, previous projects and will also typically have had time to prepare individually before meeting as a group. At the same time, it allows us to elicit a measure of individual inputs into the team. In Study 1, each team member worked alone at a separate desk facing the wall for the first five minutes. In Study 2, we increased the preparation time to 8 minutes.

For the remaining time of Part 2, participants worked collectively as a group. We refer to this as the *group phase*. Groups were free to choose whatever process they wished to arrive at a solution. As in the individual preparation phase, a worksheet was provided with room for the final answer as well as for the steps taken in deriving this final answer. The size of the reward was only tied to the final answer (in Study 2, we stressed that only answers that were well documented would be paid out). In Study 1, the team was seated at a large table in the middle of the room. The team filled in their answer on a sheet of paper. In Study 2, the team communicated anonymously via chat and each team member was able to share a copy of their individual answer sheet as well as copy steps from the individual to the group answer sheet, if they wished to do so. In Study 2, we implemented speed pay in the group phase to create opportunity costs. A team submitting their answer early could earn a small extra amount of money.

[12] An example of a filled-in worksheet from Study 1 is provided in the Supplementary Appendix.

We are interested in the performance of teams when individual rewards depend on a peer evaluation versus teams that are only incentivized on a group level. We thus implement two treatments, BASELINE and PEER EVAL.[13] In both treatments, group earnings depended on the quality of the team guess and were calculated employing the same error-band incentive scheme as in the individual ability task but now with a maximal reward of 35 euros. In BASELINE, the group earnings were split among the group members independently of individual performance.

In PEER EVAL, the distribution of group earnings was based on the outcome of an additional stage (announced in advance) that commenced immediately after the group phase and in which each group member privately voted for the teammate who they thought "contributed more to the solution to the problem". Voting for oneself was not possible, and participants had not yet been informed about the quality of their team guess at this point. Not having perfect performance feedback at the time of the peer evaluation is typical of the settings we are interested in. For example, in higher management the consequences of strategic decisions are only slowly revealed over time and counterfactuals may be hard to observe.[14]

The individual voted most valuable by both teammates received 50% of the total group earnings, the individual with one vote received 30%, and the individual with no vote received 20%. In the event of a tie, we awarded the shares of 50%, 30%, and 20% randomly.[15] In order to avoid that differences between PEER EVAL and BASELINE were affected by the earnings structure, we chose to implement noisy group incentives in BASELINE. Thus, the group reward was split in shares of 50%, 30%, and 20% in BASELINE as well, but these were always allocated randomly.

Subsequently, and without receiving feedback on performance or on the results of the peer evaluation in PEER EVAL, participants were randomly reassigned into new groups of three, and received a new guesstimation problem. We implemented a perfect-stranger matching scheme to

[13] In Study 1 we implemented a third treatment, FLAT, where teams were paid a flat fee for submitting an answer, independent of performance. We do not discuss the results of this treatment in this paper.

[14] Whether a new product turns out to be successful may take a few years to learn. Also, when choosing not to pursue an innovation, it is hard to know what the profits would have been in case of adoption.

[15] Apart from its simplicity we also chose this particular form of peer evaluation to limit the scope for strategic behavior and thus minimize gaming possibilities as a potential channel for the inefficiency of the peer evaluation (Edwards and Ewen, 1996). To see that the scope for strategic voting is limited, consider the decision of participant A. If B and C both voted for A, A is voted most valuable team member regardless of the own vote. If they both did not, A receives 20%, the lowest share, regardless of the own vote. Finally, if A received one vote from either B or C, A's vote determines whether A receives a share of 30% for sure or the vote results in a tie with an expected payoff of 33%. Note though that A does not know the voting decision of B and C.

ensure that group members never overlapped across guesstimations. The (re-)allocation of individuals to teams was done without attention to gender. Thus, we naturally obtain male-majority and female-majority teams (2 or 3 individuals of the same gender). In this sense, the random gender team composition can be viewed as another treatment dimension. In Study 1, there were three rounds of group guesstimations using the same order of problems across sessions. In Study 2, there were two rounds with the order of problems being randomized on the session level. Another difference between the studies was that in Study 2 we elicited individual confidence measures about the preparation-phase guesses after group members had completed the preparation phase but before the group phase had started. These consisted of beliefs about own absolute guess quality (achieved error band), relative performance (own rank within the team), and—in PEER EVAL only—the expected number of votes.[16]

In Part 3, we elicited social value orientation (based on the slider measure of Murphy et al., 2011) and participants privately completed a questionnaire designed to collect information about group atmosphere, personality traits (Big 5 index (BFI-10)), and demographics.

At the conclusion of each session, participants received information on their performance and were paid in private. On average, participants earned 21.15 euros in Study 1 and 22.35 euros, rates that met or exceeded the typical earning opportunities available to them.

The guesstimation questions for Study 1 (no internet, fact sheet and calculator provided) were:

- **Individual ability – "USA Dogs":** How many dogs are there in the United States of America? (A: 73.4 million).
- **Group round 1 – "Toothpaste":** How many liters of toothpaste are used in the United Kingdom every year? (A: 46.3 million liters).
- **Group round 2 – "Weddings":** How many weddings were there in Germany in June 2006? (A: 49,500).
- **Group round 3 – "Cycling":** What is the total distance cycled in Amsterdam per day?

[16]This was (mildly) incentivized: Participants were informed that at the end of the session one of their responses to the confidence questions would be selected at random and they would receive 1 euro in case their indicated belief was correct.

(A: 2 million km)

The guesstimation questions for Study 2 (internet search encouraged, calculator provided) were:

- **Individual ability – "ICU Beds":** How many intensive-care-unit beds were there in July 2022 (in total) in the hospitals of the 20 largest German cities? (A: 6,353)
- **Group neutral stereotype – "Labradors":** How many living dogs of the breed "Labrador Retriever" were registered in the German speaking cantons of Switzerland at the end of the year 2019? (A: 19,410)
- **Group male stereotype – "Football":** What is the total distance (in km) that the 20% oldest first-league German Bundesliga players ran during all matches in the 2021/22 season? (A: 15,584 km)

Since the gender stereotype of the question could be important when studying heterogeneous treatment effects by gender, we deliberately chose one team task that we thought of as evoking a gender-biased stereotype (Football) and one we considered to be stereotypically gender-neutral (Labradors). In Section 3.2, we show that this is indeed how the questions were perceived by our participants.

A detailed description of all outcome variables can be found in Appendix A.1. The Supplementary Appendix contains reproductions of all of the materials presented to participants, including instructions for both studies, the fact sheet, and a sample answer sheet of Study 1.

3 Results

3.1 Overview

Table 9 in Appendix A.3 summarizes the characteristics of our participants for both studies and treatments. We find no significant differences in characteristics between treatments. For the following analysis, we will make use of nonparametric statistical tests (two-sided Fisher-Pitman permutation tests) conducted at the level of statistically independent observations and regression analysis with clustering at the session level. We will begin by studying *individual performance* in the ability

tasks and the preparation phases for the team tasks. We examine treatment and gender effects. Second, we examine the impact of a peer evaluation on *group performance*, with and without controlling for other explanatory variables. Third, we consider measurements of *individual and group effort*. Fourth, we ask what *determines voting behavior* in the peer evaluation treatment. Lastly, we analyze the *chat data* and briefly examine *participants' preferences over the two reward allocation mechanisms*.

3.2 Individual performance: Treatment and gender effects

Throughout the results section, our measure of performance will be the achieved payoffs.[17] Recall that in the ability tasks, the maximum prize was 10 euros and in the team tasks it was 35 euros. For better comparability, we do not include the small additional rewards that were paid out for submitting a guesstimate early in Study 2. While the individual preparation phases for the group tasks in Part 2 of the experiment were not themselves incentivized, we will apply the same 35-euro prize to determine relevant hypothetical payoffs as a measure of individual performance.[18] If an individual or group did not write down a guess, we count this as a payoff of zero.

Figure 1 summarizes the average individual payoffs across all tasks for both treatments. Since the ability stage is identical in BASELINE and PEER EVAL, we do not expect a treatment effect here, nor do we observe one. Perhaps contrary to expectations, individual performance in the preparation phase for the group tasks appears to be, if anything, *worse* with the looming peer evaluation than without it. However, none of the differences is statistically significant at the 5-percent level.[19] Furthermore, there is no discernible treatment effect when we pool the preparation-phase data for each study (Study 1: $p = 0.143$; Study 2: $p = 0.430$) or overall ($p = 0.408$).

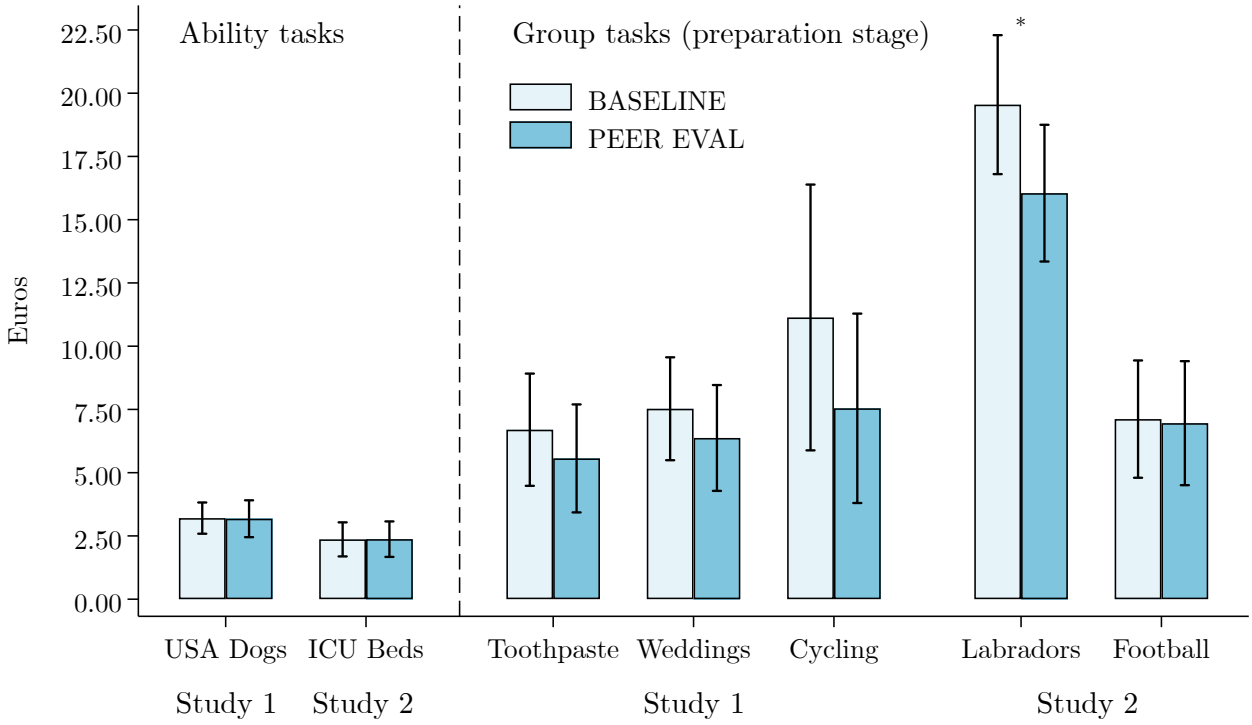
An exploratory regression analysis of whether the personality data (Big 5 index, social value

[17] An alternative way of assessing individual and group performance would be to directly use the percentage error of the submitted guesstimate. On the one hand, these data would be more sensitive to small differences in accuracy than the payoff data that are based on a step function, as explained above. On the other hand, there are several extreme outliers where guesstimates are too high by several orders of magnitude, which could bias some of our analyses. When the error data is winsorized at the 100-percent mark, we find that our results are qualitatively unaffected by which performance measure is used.

[18] As we will see, individual performance in the preparation phase is significantly related to team performance in the group phase. Since it is not possible to solve the guesstimation tasks in one's head and the sheets facilitate an efficient sharing of information with one's teammates, having a high-quality preparation sheet was an important input into the group phase.

[19] For the Labrador task we get marginal significance ($p = 0.086$).

Figure 1: Individual performance across tasks and treatments.



Individual performance measured by average achieved payoffs for each task and treatment. 95% confidence intervals are based on the means of statistically independent observations. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

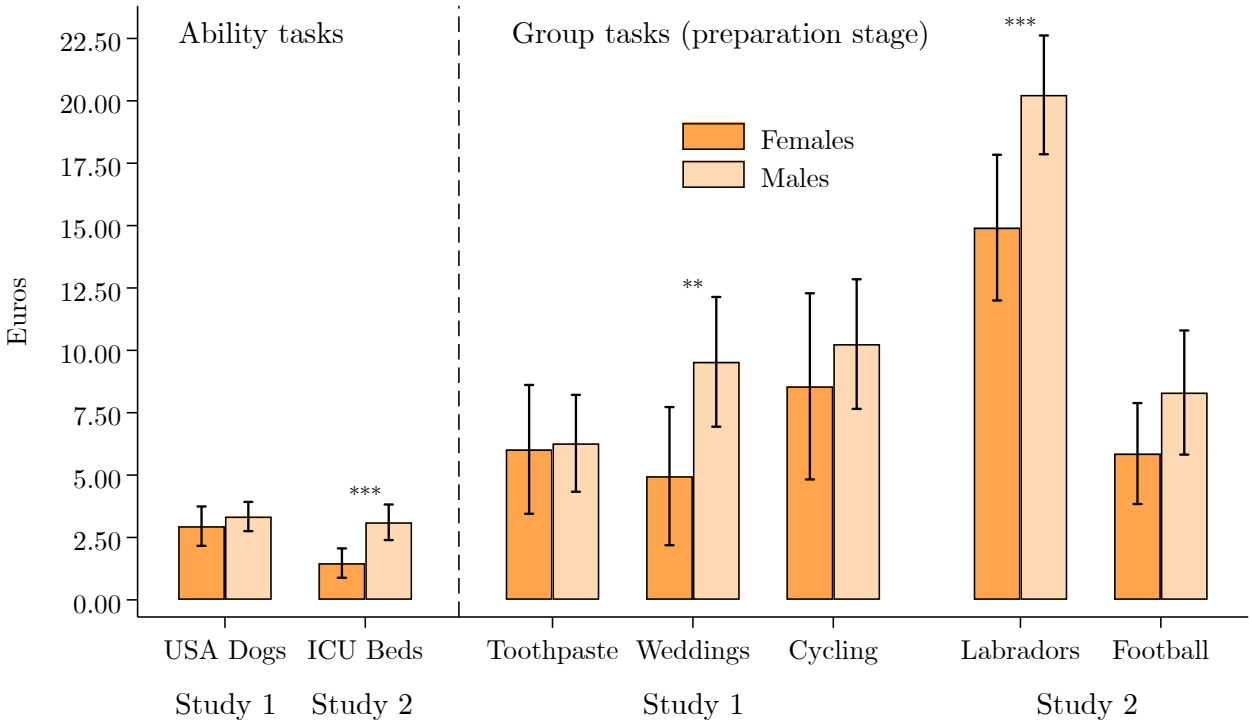
orientation), the available demographic information (age, gender, field and level of study, nationality), or self-reported previous experience with guesstimation tasks predicts differences in average individual performance does not provide much further insight (see Table 10 in the appendix).[20] By and large, these factors cannot explain the variability in individual performance, with only a few exceptions. There is a marginally significant negative age effect in the preparation phase of Study 1, and STEM-field students perform better in the ability stage of Study 2. The strongest (and for us surprising) effect is gender in Study 2: Achieved payoffs are significantly lower for women than for men, in both parts of the experiment.

To explore this issue in more detail, consider Figure 2, which again shows individual performance across tasks, but this time split by gender. As is apparent from the figure, there is considerable variability in achieved payoffs and for some tasks the null hypothesis of no gender difference cannot

[20]Table 9 in the Appendix presents summary statistics of the individual control measures by study and treatment. We have a few missing observations for the personality data (3 individuals for SVO, and another 3 for Big-5). To avoid biasing the results when we include these controls in our analyses, we use mean imputation to estimate the missing values.

be rejected. In other cases, however, the gap is extremely large, for example in excess of 100% in the ICU beds task and in the Weddings task. Overall, when we aggregate the individual performance data for each participant, we find that men earn on average about 30% more than females in Study 1 ($p = 0.053$) and about 40% more in Study 2 ($p = 0.001$).[21]

Figure 2: Individual performance across tasks and gender.



Individual performance measured by average achieved payoffs for each task and for males vs. females. 95% confidence intervals are based on the means of statistically independent observations. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In designing Study 2, we aimed to have one gender-neutral and one stereotypically-male question as group tasks. As part of the post-experimental questionnaire, we elicited participants' perceptions of whether the group tasks seemed easier for either men or women.[22] And indeed, for the Football task, responses indicate (for both male and female responders) a strong belief that men will find this task easier. In contrast, the perception of the relative difficulty of the Labrador task appears to depend on the respondents' own gender: Men anticipate this task to be gender-neutral while women

[21] The magnitudes of the differences are even slightly larger when we use the predicted average earnings based on regression analyses that include additional controls (other demographics, self-reported experience, personality data) for the comparison. With the additional controls, the gender difference remains significant in Study 2 ($p < 0.001$) but not in Study 1 ($p = 0.174$).

[22] A detailed breakdown of the results from this analysis is provided in Appendix A.2.

think it favors women. Thus, remarkably, our participants get it wrong in both cases: They fail to anticipate the strong gender effect in the Labrador task, where in fact males earn substantially more than females, and they assume that men have an advantage over women in the Football task, where in fact our data does not provide evidence for a systematic performance difference between the genders.

Interestingly, the data from our confidence measures paints a different picture. After having completed the preparation phase of the Labrador task, men are significantly more confident about their performance than women in terms of absolute performance and in terms of the probability of winning the voting stage ($p = 0.007$ for both; $p = 0.131$ for relative-performance confidence). But after the preparation phase of the Football task, there is no clear gender difference in confidence ($p = 1.000$, $p = 0.122$, $p = 0.275$ for the absolute, relative, and voting confidence measure, respectively). Thus, when men and women are asked about their beliefs regarding their *individual* performances rather than what they think about the performance of men vs. women in general, they seem collectively better able to give a realistic sense of the relative level of difficulty.

Finally, we examine whether peer evaluation affects men and women differently. We indeed find this to be the case in Study 1. For men, the average individual performance across tasks (excluding the ability stage which cannot plausibly be affected by treatment) is virtually the same across treatments (they earn 7 cents less in the peer evaluation treatment; $p = 0.991$), but for women, there is a dramatic change: Their payoff drops by 51% in PEER EVAL relative to BASELINE ($p = 0.009$). This is in part caused by women's guesses becoming worse and in part by a much larger number of *missing* guesses from women in the preparation phase of the peer evaluation treatment.[23] We will discuss underlying reasons for this finding when we examine our measures of effort and motivation in Section 3.4.

These results from Study 1—lower performance and fewer guesses from women who face the prospect of being peer-evaluated—led us to hypothesize that females might experience some form of 'choking under pressure' under peer evaluation (Ariely et al., 2009, Bracha and Fershtman, 2013).

[23]Comparing BASELINE and PEER EVAL, the proportion of missing female guesses increases from 13% to 31% ($p = 0.017$). However, the performance of women becomes worse even when we focus on payoffs from submissions without missing guesses ($p = 0.043$). For comparison, the proportion of missing guesses from men remains virtually the same (19% in both treatments).

Exploring this idea partly motivated our second study. However, as it turns out, we are unable to replicate the gender-specific changes in performance in Study 2. In both the Labradors and the Football task, *men* do worse in PEER EVAL relative to BASELINE (-21% for Labradors and -22% for Football), although only the Labradors case is marginally significant ($p = 0.063$). For women we find no statistically significant differences between treatments. In the Labradors task, their performance is a little lower under peer evaluation than in the baseline (-9%), but in the (stereotypically male-favoring) Football task their performance even improves with peer evaluation ($+20\%$).[24]

3.3 Group performance with and without peer evaluation

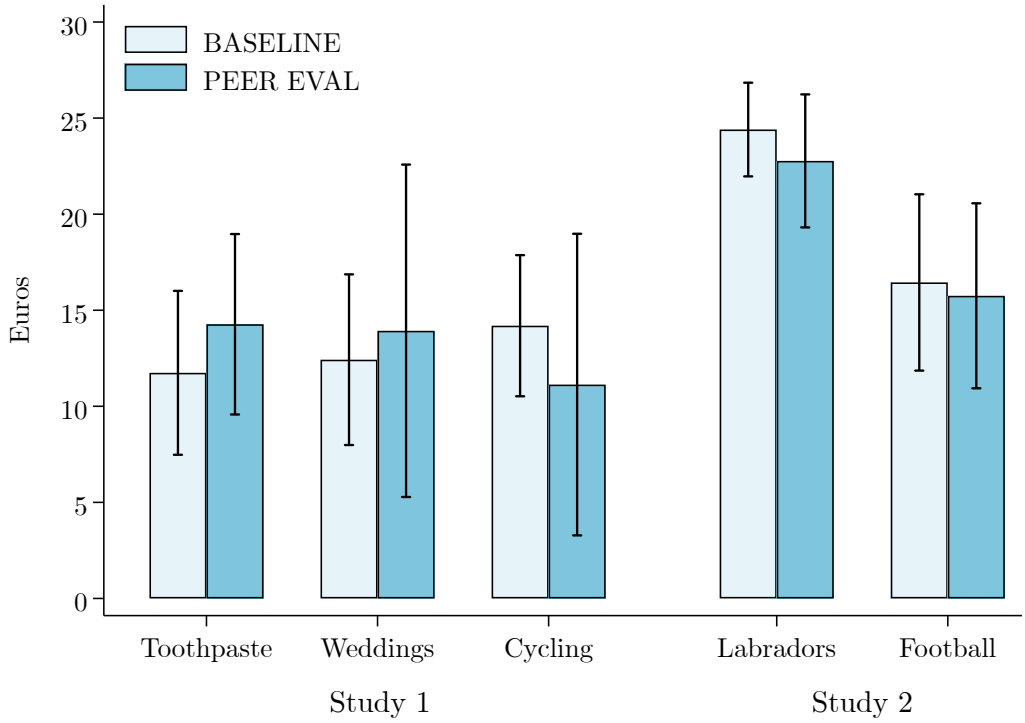
We now examine how well *groups* perform and whether the treatment change affects their performance. Figure 3 displays the average achieved payoffs for teams in part 2 of the experiment. Comparing the team performance to the average within-team individual performance in the preparation phase, we find that working together leads to large and significant improvements in the quality of guesses (Study 1: $+73\%$; Study 2: $+52\%$; $p < 0.001$ in both). However, team performance does not keep up with the *best* within-team performance in the preparation phase (Study 1: -18%, $p = 0.007$; Study 2: -13%, $p = 0.013$). Thus, while teamwork clearly adds value to what individuals can achieve on their own on average, teams do not appear to be able to easily recognize (let alone improve upon) the best individual guess. At the same time, the best individual guess is not irrelevant: Spearman's rank correlation coefficient is $\rho = 0.54$ in Study 1 and $\rho = 0.44$ in Study 2, suggesting that better individual guesses do translate into better group guesses.[25]

As is apparent from the figure, we again find no treatment effect. Overall, team payoffs in PEER EVAL are 3.9% higher compared to BASELINE in Study 1, and 2.7% lower in Study 2. Neither of these differences is statistically significant ($p = 0.749$ and $p = 0.747$ for Study 1 and Study 2, respectively).

[24] Note also that we find no statistically significant treatment differences in the confidence and belief data for women in Study 2. Likewise, the share of missing preparation-phase guesses in Study 2 does not appear to be affected by gender or treatment. At around 6% it is also much lower than in Study 1, although this is not too surprising given the computerized setting with a countdown timer displayed on screen.

[25] Most groups do not simply adopt one of the preparation phase guesses as their group guess. This only happens in around 10% of groups in both studies. Typically groups discuss the guesstimation step by step.

Figure 3: Group performance across tasks and treatments.



Group performance measured by average achieved payoffs for each task and treatment. 95% confidence intervals are based on the means of statistically independent observations. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1 reports the results of a regression analysis of group payoffs on the treatment variable controlling for various other factors. Columns 1-3 report the results for Study 1, columns 4-6 for Study 2, and column 7 is based on the pooled data. Models 1 and 4 essentially replicate the results from the nonparametric analysis, now with controls for team members' average performance in the ability stage and for group-level aggregates of their individual characteristics (see the notes underneath the table).

In models 2 and 5, motivated by the treatment effect we observed in the preparation-phase performance of women (see Section 3.2), we interact the treatment variable with an indicator for teams with two or three female members. The significantly negative coefficient on the majority-female teams/treatment interaction in Study 1 suggests that the negative effect from the preparation phase does carry over to the group phase for the majority-female teams.[26] However, just like in the

[26]More precisely, the treatment effect on performance is significantly worse for majority-female groups than for majority-male ones. The absolute treatment effect for majority-female groups is also negative ($p = 0.077$; Wald test on the sum of coefficients for PEER EVAL and PEER EVAL × MajFem).

Table 1: Determinants of group performance

	Study 1			Study 2			Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PEER EVAL	-0.270	2.077	3.192	1.969	1.908	2.387	1.355
	(1.699)	(1.819)	(1.861)	(2.179)	(2.173)	(2.102)	(1.255)
Average ability pay	0.416	0.339	0.355	0.238	0.239	0.0323	0.164
	(0.641)	(0.615)	(0.350)	(0.405)	(0.406)	(0.384)	(0.241)
PEER EVAL × MajFem		-7.147**	-3.280*		0.154	0.767	-0.986
		(2.446)	(1.650)		(4.225)	(3.258)	(1.837)
Max prep-phase pay			0.412***			0.230*	0.355***
			(0.105)			(0.114)	(0.0776)
Median prep-phase pay			0.143			0.162	0.148*
			(0.138)			(0.0978)	(0.0783)
Min prep-phase pay			0.397**			-0.0320	0.130
			(0.165)			(0.0930)	(0.0774)
Missing prep-phase guesses			1.426			-3.499	-0.0657
			(1.322)			(3.023)	(1.121)
Study 2							6.724**
							(2.474)
Additional covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	171	171	171	134	134	134	305
Clusters	11	11	11	20	20	20	31

Dependent variable: Achieved group payoffs. The table reports OLS estimates with robust standard errors clustered on the session level and corrected using wild bootstrap (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). Additional covariates consist of indicator variables for the number of females in the group, of the average age of group members, the per-group number of participants with the local nationality (Dutch in Study 1, Germans in Study 2), of economics and business students, of STEM-field students, of bachelor students in their first or second year, and of members who report having experience with guesstimations, group average scores for each of the Big5 items and the number of team members categorized as prosocial/altruistic (social value orientation), plus dummies for task and round number for Study 2.

case of the individual-performance data, we cannot reproduce this result in Study 2.

For models 3 and 6, we consider measurements of the group members' individual performances in the preparation phase in the form of the best, median, and worst guess payoff, and also add the number of missing preparation-phase guesses, which—as we have seen—increase for females in Study 1 under peer evaluation. The results confirm that individual performance is an important contributor to team performance. Furthermore, including these controls halves the Study 1 coefficient for the interaction term and renders the effect of the peer evaluation for female-majority groups insignificant.[27]

Taken together, these findings suggests that the negative treatment effect for majority-female teams in Study 1 is in essence driven by underperformance in the preparation phase. In Study 2, there is no gender-treatment effect to begin with and this does not change when we consider each guesstimation problem on its own. Neither in the Football task, which participants believe to favor men, nor in the Labradors task, which they expect to be more gender-neutral or even female-favoring, do we find any gender-specific treatment effects at the group level, whether we do or do not include the additional group-level controls.

3.4 Effort

3.4.1 Individual effort: The preparation phase

Our results on individual and team performance leave us with a puzzle: Why is the introduction of individual incentives ineffective? Are group incentives or a possible intrinsic enjoyment of working on guesstimation tasks already sufficient to motivate participants to the maximum? To shed further light on the effects of peer evaluations, we now turn to measures of effort. We begin our analysis with the preparation phase of the group tasks. Table 2 provides an overview of our measures of output that the individuals produced on their own before coming together as a team. The measures are listed by treatment and separately for each study. In Appendix A.3 we pool the variables that we obtained in both studies for a joint analysis, which we also conduct for males and females separately (Table 11).

[27] Wald test: $p = 0.948$.

Table 2: Individual effort measures (preparation phase of the group tasks)

		BASELINE	PEER EVAL
Prep unfinished	Study 1	0.18	0.27
Prep timeout	Study 2	0.29	0.37
Prep work time	Study 2	7.33	7.52*
Prep steps	Study 1	4.83	4.80
	Study 2	3.80	4.19
Prep characters (no URLs)	Study 2	223.61	255.26*
Prep # URLs	Study 2	1.16	1.35
Prep Different	Study 1	0.10	0.11
Prep Different URL	Study 2	18.28	16.17**

The table reports group means. Asterisks (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$) refer to the treatment comparison (PEER EVAL vs. BASELINE). The p-values for *Prep finished*, *Prep timeout*, and *Prep different* are calculated from probit regressions, and the other ones from OLS regressions. All specifications cluster standard errors at the session level corrected by using wild bootstrap (STATA boottest) and contain the full set of individual control variables shown in Table 9. For Study 2 we also include a question/round indicator. For *Prep different* in Study 2 we include only sheets that have at least one URL and we control for the number of URLs in addition.

We know from the previous section that females in Study 1 performed significantly worse under peer evaluation, partly due to them failing to write down a final answer for their preparation-phase guesstimate (see Footnote 23). Sheets without a final answer may be an indication for individuals exerting effort but then running out of time. An alternative interpretation is that time is not the issue but individuals shy away from committing to a definitive final answer on which they can be judged by their teammates. To better understand under which circumstances missing guesses occur, we tasked a research assistant to (subjectively) code whether the step-based reasoning on an answer sheet seems—irrespective of whether or not there is a final guess—to have been completed or not (variable *Prep unfinished* in the table). We find that the correlation between missing guess and the answer sheet being coded as unfinished is 0.89, corroborating that missing guesses seem to be predominantly due to unfinished preparation work.

In Study 2 we measured work time directly and find that participants spend more time on their preparation in PEER EVAL than in BASELINE (the change is marginally significant). The Study-2 data also contains information on *timeouts*, which could be interpreted as signs of effort. Timeouts occur when the on-screen countdown reaches zero and the page form is submitted automatically by

the computer. They do not necessarily lead to missing guesses as the software tries to retain all form data that was entered on the page. As shown in Table 2, the number of timeouts is higher in the peer evaluation treatment but not significantly so. Considering timeouts in Study 2 and unfinished methods in Study 1 as both measuring timeouts, we find a marginally significant increase in the occurrence of timeouts in the pooled data (Table 11). Disaggregating by gender, we find that this increase is driven by females.

Yet another measure of effort is the length of the filled-in answer sheets. Counting the steps taken by each individual to derive a preparation-phase guesstimate, we find overall no systematic difference between treatments in either study, neither if we pool the data. For Study 2 we also have readily available data on the number of characters typed as an alternative measure of answer sheet length. Here we do see a positive and significant effect of the peer evaluation. We also consider the number of URLs inserted into the answer sheet but do not find any treatment effects.

Finally, we consider measures of the uniqueness of individual approaches as a proxy for 'creative effort'. As explained in more detail in Appendix A.1, in Study 1 we had research assistants code the uniqueness of the steps (variable *Prep different* in the table). In Study 2, we instead use the uniqueness of the URLs on the answer sheet (see again Appendix A.1 for more details), and find a significant increase in the uniqueness of URLs in PEER EVAL compared to BASELINE (variable *Prep different URL*). We do not find any effects in the pooled data, where we combine both measures.

Overall, our results provide at least weak evidence that peer evaluation encourages participants to exert more effort in the preparation phases of the group tasks. The evidence is a little stronger for females.

3.4.2 Team effort: The group phase

To examine the effects of peer evaluation on team effort, consider Table 3, which displays the means of a range of measures by treatment and study. In Table 12 in Appendix A.3 we report on regressions where we pool the variables we obtained for both studies. We also consider differential effects by team composition.

Our first result is that in both studies, groups work significantly *longer* in PEER EVAL than

Table 3: Measures of effort during the group phase

		BASELINE	PEER EVAL
Work time	Study 1	8.37	8.83*
	Study 2	11.47	12.95***
Timeout	Study 1	0.36	0.49
	Study 2	0.30	0.46*
Sheet characters (no URLs)	Study 2	231.18	248.87
Number URLs	Study 2	1.06	1.04
Steps	Study 1	4.51	4.64
	Study 2	4.05	4.10
Different	Study 1	0.04	0.09
	Study 2	7.32	5.70
Unrelated	Study 1	0.49	0.49
Number of speaking turns	Study 1	10.45	11.89
Chat messages	Study 2	37.14	47.22
Chat words	Study 2	254.18	355.47**
Methods shared	Study 1	1.20	1.08
	Study 2	2.59	2.62

The table reports group means. Asterisks (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$) refer to the treatment comparison (PEER EVAL vs. BASELINE). The p-values for *Timeout*, *Different* in Study 1, and *Unrelated* are calculated from probit regressions, and the other ones from OLS regressions. All specifications cluster standard errors at the session level corrected by using wild bootstrap (STATA boottest) and contain the full set of group control variables shown in the table notes of Table 1. For Study 2 we also include a question/round indicator. For *Different* in Study 2 we include only sheets that have at least one URL and we control for the number of URLs in addition.

in BASELINE, a result also visible in the pooled analysis in Table 12. In the data from Study 2, we also find an increase in timeouts, reflecting the potential negative effect of peer evaluations we already noted for females in Study 1[28]. While the increase is not significant for the Study-1 subsample, timeouts significantly increase under peer evaluation ($p < 0.05$) when combining both studies.

Interestingly, the longer work time in the peer evaluation treatment does not translate into increases in the number of steps or (in Study 2) increases in the number of URLs or of characters typed on the group answer sheets. We also do not observe a significant increase in our uniqueness

[28]Groups who do not encounter a timeout earn on average more than 5 euros more than groups who do not run out of time, even when missing team guesses, which imply a payoff of zero, are disregarded ($p = 0.063$).

measurements. Instead, based on the recorded face-to-face conversations in Study 1 and on the chat-messages data from Study 2, we find that team members appear to *communicate* more when the group phase is followed by a peer evaluation stage. This increase is statistically significant for the number of words used in the Study-2 text messages. The number of speaking turns and the number of chat messages increase as well, but these differences between the treatments do not reach statistical significance. Pooling both across studies as a joint measure of *communication turns*, we find a significant increase under peer evaluation in Table 12. However, we do not observe that increased communication is accompanied by more sharing of preparation-phase information.

Disaggregating the results of our pooled analysis by gender composition of the team (Table 12), we find that the increase in work time and timeouts is mainly driven by male-majority teams. Female-majority teams exhibit significantly more communication turns and a significantly higher share of unique team answer sheets under peer evaluation.

How can we interpret the increase in work time and communication when peer evaluation is introduced? Neither of these are unambiguous measures of "productive" effort. While, for example, working longer on the task may be an indication of a larger investment into making the group guess better, it may equally well just be an indication of individuals increasing their efforts to stand out within the group. Trying to impress the other group members may indirectly also turn out to be productive for the group, but this does not have to be the case. Instead, it may even deteriorate information exchange as individuals trying to impress may be less open to listen to their teammates' views or may be exaggerating the confidence in their own answer, thus distorting group decisions. In the next section we study participants' reported motivation and group perception to shed further light on this question.

3.4.3 Motivation

We conclude our analysis of effort by briefly looking at some post-group phase questionnaire data. Table 4 reports the mean answers of participants regarding their motivation and their perception of the group atmosphere (in each task) by treatment and study. As the table shows, responses to our questions change between treatments in both studies. Participants agree significantly more with

the statement "I wanted to make a good impression on my group members" in the peer evaluation treatment. They also perceived the atmosphere to be more competitive in PEER EVAL than in BASELINE. While this may or may not encourage effort, in Study 1 there is also more agreement with the statement that "I felt that others dominated the discussion in unproductive ways" under peer evaluation. This points towards a potential performance-reducing effect of peer evaluation especially for face-to-face work teams.[29] In Study 2, we also asked about group and individual effort directly. We find that participants are more inclined to report that they personally did their best and that their team did its best when working under peer-evaluation incentives. We do not observe significant changes in self-reported stress level in Study 2, nor do we find gender differences in the PEER EVAL responses.[30]

Table 4: Perceptions of group environment by treatment

		BASELINE	PEER EVAL
"I wanted to make a good impression on my group members"	Study 1	3.64	4.12***
	Study 2	3.52	3.98***
"The atmosphere in the group was helpful"	Study 1	4.09	4.16
	Study 2	3.93	4.05
"The atmosphere in the group was competitive"	Study 1	2.29	2.71***
	Study 2	2.11	2.48*
"I felt that others dominated the discussion in unproductive ways"	Study 1	1.79	2.05***
	Study 2	1.86	1.95
"I felt that everyone had an opportunity to voice their ideas in a fair way"	Study 1	4.43	4.38
"I worked hard to come up with the best possible answer."	Study 2	4.23	4.46***
"All in all, my team did its best to come up with a good answer."	Study 2	4.18	4.27*
"During my work on this task, I felt stressed (in a negative way)."	Study 2	3.32	3.47

The table reports group means. Answers to the survey questions range from 1 fully disagree - 5 fully agree. Asterisks (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$) refer to the treatment comparison (PEER EVAL vs. BASELINE). The p-values are calculated from ordered logistic regressions. All specifications cluster standard errors at the session level corrected by using wild bootstrap (STATA boottest) and contain the full set of individual control variables shown in Table 9. For Study 2 we also include a question/round indicator.

[29] Combining the data from both studies, the effects for "impression", "competitive" and "dominated" stay strongly significant ($p < 0.01$) (not reported).

[30] We run ordered logistic regressions where we include a gender/treatment interaction, which is never significant.

3.5 Determinants of the peer evaluation

We now turn to the determinants of success in the peer evaluation stage of the PEER EVAL treatment to see which types of behavior were rewarded by group members. Table 5 shows how often the best, median, and worst guess ended up winning the peer evaluation as well as how many peer evaluations ended in a tie. We see that the best individual preparation-phase guess was *not* rewarded in the peer evaluation. In Study 1, both the median guesser and the worst guesser had a higher likelihood of winning the peer evaluation.[31] In Study 2, the best guesser had a higher chance of winning than the worst guesser but not than the median guesser. Hence, factors other than objective guess accuracy must have determined voting behavior. Only 15% [9%] of peer evaluations ended in a tie in Study 1 [Study 2]—thus, this was a relatively unlikely outcome.

Table 5: Peer evaluation winner by preparation guess rank

	Best guess	Median guess	Worst guess	Tie
Study 1	22%	31%	32%	15%
Study 2	37%	37%	16%	9%

To get a better sense of the reasoning behind the individual votes, we asked participants at the end of the experiment whether they voted "strategically" in any of the peer evaluations. 15% [14%] of individuals in Study 1 [Study 2] answered this question in the affirmative. When asked about the underlying reason for their vote in Study 1, 12% of participants stated that they voted mostly or nearly all on 'manner', 33% on 'equally manner and content', while the majority of 54% voted 'mostly or nearly all content'. We also asked participants to state their reason in an open text field. A research assistant coded the participants' answers into 6 categories, where multiple categories per answer were possible. We find that in both studies the most common category refers to perceived performance in the preparation-phase sheets[32] (83% in Study 1, 45% in Study 2), while the second largest category refers to communication (30% in Study 1, 40% in Study 2). Interestingly, we see that in Study 1 the preparation-phase category was much more prevalent than the communication

[31] Note that worst guesses are often missing guesses.

[32] This could be about certain content that they brought into the group phase from their answer sheet, the preparation guess itself, the URLs, or anything else related to their input from the preparation stage.

category, while they are nearly equally prevalent in Study 2. Furthermore, 23% (respectively 6%) of vote reasons refer to personality, 8% (respectively 33%) to effort, and 4% (respectively 7%) to similarity with the own approach. 7% and 17% bring up leadership. Thus, while participants state that the quality of the preparation phase answer sheet mattered, other aspects such as behavior during the group phase were important for their peer evaluation vote as well.

In order to understand these other aspects better, we now examine the determinants of the peer evaluation using regression analyses. Table 6 presents the results of an OLS regression on the chance of an individual being voted winner (i.e., receiving a vote from both peers in the peer evaluation). Columns (1) and (2) present coefficients for Study 1 and Study 2 separately, while in column (3) we pool the data for both studies. In column (4) we restrict the analysis to groups where no participant stated that they voted strategically. While the scope for strategic behavior to increase one's payoff was very limited in our setting, it is still interesting to see whether individuals who report having voted non-strategically are more inclined to reward the best individual guess.

Several things are interesting to note. First, confirming our impression from Table 5, having had the objectively best answer in the group is never significantly positively related to being voted most valuable team member. Conversely, not having an individual answer at all is also not significantly related to the chance of winning the vote.

Other individual outcome measures do predict success in the peer evaluation. In Study 1, the number of steps on the individual answer sheet (*Steps*), having applied a "different" method (as coded by research assistants, see above), and the share of speaking turns (*Communication share*) are positively and significantly related to the probability of winning the peer evaluation. Interestingly, being male is not an additional predictor of success in the peer evaluation in Study 1, when controlling for our effort measures and covariates. Thus, the gender difference in preparation-phase performance cannot be explained by females rationally expecting to be treated differently from males. In Study 2, the only significant predictor of peer evaluation success is the individual share of messages written in the group chat (*Communication share*).

Pooling the data from both studies, we find both *Steps* and *Communication share* to matter.[33]

[33]Running the regression in column (3) for males and females separately to study whether the criteria differed by gender, similar results are obtained. *Communication share* is a significant predictor for both genders. *Steps* is

Table 6: Determinants of winning the voting stage in PEER EVAL

	Study 1 Full Sample	Study 2 Full Sample	Pooled Full Sample	Pooled Non-strategic
Best guess	-0.136	0.0189	-0.0373	-0.0115
	(0.0753)	(0.0701)	(0.0485)	(0.0512)
Missing guess	0.0173	0.132	0.0787	0.0642
	(0.0599)	(0.0857)	(0.0496)	(0.0609)
Steps	0.0509*	0.0293	0.0430**	0.0280
	(0.0217)	(0.0205)	(0.0153)	(0.0207)
Prep different	0.136*	-0.0555	0.0288	0.172**
	(0.0522)	(0.0846)	(0.0566)	(0.0689)
Filled in sheet	-0.0246	0.0532	0.0227	0.0866
	(0.0641)	(0.114)	(0.0645)	(0.0749)
Communication share	0.402**	0.980**	0.831***	0.885***
	(0.138)	(0.326)	(0.189)	(0.257)
Shared method	0.0265	0.101	0.0572	0.139***
	(0.0392)	(0.0974)	(0.0412)	(0.0404)
Male	-0.0774	0.0144	-0.0270	0.116
	(0.0883)	(0.0881)	(0.0693)	(0.0672)
Additional covariates	Yes	Yes	Yes	Yes
Observations	225	204	429	261
Clusters	5	10	15	15

Note: The table reports OLS regression estimates of the determinants of the PEER EVAL vote for best team member. Columns (1) and (2) report results for Study 1 and Study 2 respectively, while columns (3) and (4) consider the pooled sample. The sample for Study 1 consists of all participants who received the PEER EVAL treatment minus 9 observations due to technical problems with the video. The set of individual control variables shown in Table 9 is included in all specifications. Except for Study 1, we also include a question/round indicator. Robust standard errors clustered by session are reported in parentheses. They are corrected using wild bootstrap for columns (1) and (2). Asterisk indicate significance at the 10/5/1 percent level. *Different method* in Study 2 is an indicator for the group member with the, on average, most unique URLs on their preparation sheet if their uniqueness score is in the bottom 25% of all preparation uniqueness scores. *Communication share* is the share of speaking turns in Study 1 and the share of messages in the chat in Study 2. *Filled in sheet* in Study 2 is 1 for the team member that had the most edits in the group.

Finally, for the subsample of groups that did not vote strategically, we find, in addition, a positive and significant relationship of peer evaluation success with whether an individual shared their own answer sheet with the group (*Presented method*) and whether a "different" method was applied.

Since we did not provide any feedback on the quality of the group answer before the evaluation, judging the accuracy of an individual guess may obviously have been difficult for participants. Our results are consistent with participants considering the individual solution steps of each group member and how they were presented and communicated as a proxy for their group contribution in their vote.[34] Evidence supporting the complexity of judging an answer can be found by studying individual confidence in the preparation phase answer relative to the actual quality of the answer, which we summarize in Table 7. Interestingly, in PEER EVAL there is little difference between the expectations of the best and worst guesser.[35] Overall, it seems that participants found it hard to judge the absolute and relative quality of their answer and participants with worse preparation guesses tended to overestimate the quality of their guess. While this measure was elicited before entering the group phase, it may be a sign that judging the quality of an answer was difficult and participants needed to rely on other cues. This is also consistent with our finding that the best individual guess has a significantly higher quality than the group guess (see Section 3.3).

3.6 Chat analysis

In this subsection, we utilize the text-messages data from Study 2.[36] Table 6 already showed that the share of speaking turns and the share of words in the chat are strong predictors of success in the voting stage of the PEER EVAL treatment in both studies. Table 3 showed that the total number of words in the chat increased significantly under a peer evaluation in Study 2.

One reason for the increased intensity of communication may be that individuals work harder

significant for males only.

[34] Job aspirants preparing for guesstimation questions in an assessment center are often given the advice that the steps undertaken are at least as important as their final answer. For example, in the article "How to Solve Google's Crazy Open-Ended Interview Questions" on wired.com by Daniel Levitin (https://www.wired.com/2014/08/how-to-solve-crazy-open-ended-google-interview-questions/), the reader is reminded: "And remember, the final number is not the point—the thought process, the set of assumptions and deliberations, is the answer." While a company will ultimately be interested in their bottom line, i.e., the correctness of the final guess, the quality and creativity of the answer steps seem to be regarded as important indicators of this in practice.

[35] This is especially pronounced in the Football task, which was the more difficult of the two.

[36] Appendix A.4 contains a word cloud per treatment to illustrate the most frequent words used in the group chat.

Table 7: Confidence and actual performance in Study 2

		Best guess	Median guess	Worst guess
Peer Evaluation	Actual Error Bin	37	60	75
	Expected Error Bin	39	40	44
	Expected Rank	2.0	2.0	2.1
	Expected # Votes	1.0	0.9	0.8
Group Incentives	Actual Error Bin	36	56	74
	Expected Error Bin	36	46	52
	Expected Rank	2.0	2.1	2.3

and thus spend longer discussing the answer steps. An alternative would be that individuals spend time promoting themselves and their answer in order to increase their chance of winning the peer evaluation. To test for these channels, we study the occurrence of certain groups of words in the chat of each group member by treatment. Table 8 summarizes which words we grouped together under which topic.[37] The first two groups, *Labradors topic* and *Football topic*, collect words related to the task at hand, and thus measure engagement with the guesstimation question. *URLs* collects the number of URLs in the chat and thus measures the sharing of information sourced from the internet. All of these could be seen as measures of productive work effort.[38] *Positive* and *Negative* mean to capture the tone of communication. *Leadership* intends to capture leadership initiative by an individual. Finally, *Time Pressure* intends to capture instances where teams appear to be running out of time.

The last two columns of Table 8 reveal an immediate first result: *All* word groups are more frequent in PEER EVAL. To study the robustness of this, we regress the number of words of each group on treatment to study the effect of the peer evaluation on the absolute and relative use of these word groups. Table 13 in Appendix A.3 summarizes our results for the absolute number of words in each category. We find that PEER EVAL participants use significantly more words from the groups of *Labradors topic*, *Positive*, and *Leadership* than participants in BASELINE. Table 14 includes an additional control for the *total* number of words used in the chat and thus looks at

[37] To find these words, we counted all instances of the strings shown in the table in each individual's chat communication. Strings were case insensitive and may be embedded in a longer word (e.g., "dog" will also count "dogs").

[38] Though one could surely come up with an argument how even on-topic conversations may be detrimental to performance, for example by presenting unnecessarily detailed information to showcase one's knowledge.

Table 8: Word groups

	Strings	BASELINE	PEER EVAL
Labradors topic	"dog", "switzerland", "labrador", "german", "speaking", "canton", "retriever"	4.3	6.7
Football topic	"football", "player", "distance", "german", "oldest", "bundesliga", "season"	4.7	5.4
URLs	"http"	0.3	0.5
Positive	"thank", "yes","right", "good", "agree", "yeah"	1.8	2.9
Negative	"wrong", "mistake", "dislike", "disagree", "sorry"	0.1	0.2
Leadership	"let's", "let us", "we should", "we can"	0.4	0.9
Time Pressure	"need to submit", "no time", "of time", "hurry", "quick"	0.5	1.0

the relative number of words used from a certain group. We find that the PEER EVAL treatment results in participants being significantly more inclined to use words from the group of *Leadership* even when we control for the total number of words.

Previous literature, such as Hardt et al. (2023), find that males communicate more than females when working as a group. We find gender effects in line with this literature. In particular, we see that males use significantly more of the topical words (the *Football Topic* and *Labradors Topic* categories), an indication that they work harder and try to initiate leadership. Table 14 shows that for the Labradors task this is even true when controlling for the absolute number of words. We do not find any gender-treatment effects.

To summarize, our exploratory chat data analysis suggests that groups whose members know that they will have to evaluate each other work harder on the Labradors question but not on the (very difficult) Football question. It also supports the finding that the peer evaluation increases efforts to lead the group and to create a positive atmosphere.

3.7 Preferences for peer evaluation

In Study 2 we asked participants at the end of the session which reward-sharing scheme they would prefer, the one that they had just experienced (in their treatment, BASELINE or PEER EVAL) or an alternative, which we briefly described (in fact, the respective other treatment). Overall,

13% of participants were neutral, 50% preferred PEER EVAL, and 37% the BASELINE procedure. Interestingly, stated preferences differed a lot by assigned treatment: In the PEER EVAL treatment, 69% prefer the peer evaluation procedure and 19% prefer a random procedure. In the BASELINE treatment, only 29% prefer peer evaluation, while 57% indicate a preference for the scheme they had seen in their session. The share of participants who were neutral on this question was similar across treatments (13% in PEER EVAL vs. 14% in BASELINE). Other significant determinants of a preference for peer evaluation are being altruistic/pro-social, studying Economics or Business, and the payoff in the ability guesstimation. Table 15 in Appendix A.3 present the results of an ordered logistic regression. While peer evaluation was the more popular incentive scheme overall, the incentive scheme in place always received majority support.

4 Conclusions

The age of the lone inventor or problem-solver is long since past. With the rise of specialization and the exponential growth in knowledge, the renaissance man is extinct. Most complex problem solving is done in groups. We sought to emulate this industry practice by studying group performance in complex tasks. Incentivizing such groups is difficult as individual performance is hard to observe. Peer evaluations may offer such a measure of individual contribution to the group. In this paper, we compare performance under pay that depends only on group performance with an incentive scheme that in addition ties individual pay to the outcome of a peer evaluation. We do so in two different studies. Our first study considers a traditional in-person setting of face-to-face communication where groups solved the task seated around a table using pen and paper. Our second study represents a more anonymous online interaction, where group members were seated at computer terminals, and could communicate only through chat.

Overall, we find no significant effect of peer evaluations on group performance in either setting. Nevertheless, groups behave differently under a peer evaluation. Participants report higher motivation and groups work longer and communicate more in the group phase. We also observe an increase in timeouts and unfinished preparation sheets. Looking more closely at the incentives induced by peer evaluation, we find that the mechanism is imperfect: Success in the peer evaluation

is not related to individual performance as measured by the preparation-phase guess. Instead, it is positively related with easily observable proxies of effort, such as speaking turns and number of steps. At the same time, individual guess quality is an important predictor of team performance. Our findings point towards two potentially performance-reducing effects of peer evaluations: (1) while participants state that they worked harder, they also state that they tried harder to impress the other team members, which may or may not lead to better performance, and (2) higher effort (whether productive or aimed at winning the peer evaluation) makes timeouts and unfinished work more likely, with potentially detrimental effects on performance. Overall, our results do not support the use of peer evaluations in settings with complex tasks and limited performance feedback, especially when work needs to be done under time pressure.

In an ex-post analysis of the data of Study 1, we find a differential gender effect in the response to peer evaluation. When facing peer evaluation, females underperform in the preparation phase and this translates into worse performance of female-majority teams. Male performance, in contrast, is not affected by peer evaluation. However, we do not replicate this negative effect on females in Study 2. While Study 1 and Study 2 differ in several respects, one important aspect is that participants were given a more generous time budget and had a countdown timer on their screen. This significantly decreases unfinished answers in the preparation phase. At the same time, we find an increase in timeouts in the group phase as a response to peer evaluation. Examining the role of time pressure for the effectiveness of peer evaluations, and whether it affects males and females differently, seems an interesting avenue for further research to us. Our study does not offer conclusive evidence of a differential gender effect of peer evaluations. We leave it for future research to more carefully study whether modern approaches to performance evaluation systems based on peer evaluation may be contributing to a gender gap on the labor market.

Finally, not all peer evaluations are tied to performance pay. Instead, they may be used as a mechanism to provide feedback to employees and increase future productivity (Villeval, 2020). In our experimental design we switched off the dynamic effects of peer feedback to focus on the incentive effects. We look forward to future research to better understand whether peer evaluations are performance enhancing through providing peer feedback when teams interact repeatedly.

A Appendix

A.1 Description of variables

Main performance measure

Group and individual payoff (0-35 euro) calculated using the incentive scheme and the answer to the guesstimation.

Group and individual preparation sheet characteristics

Aside from the final guesstimation answer, we collected the following measures of effort and group process from the individual and group answer sheets:

- The number of steps taken to arrive at the answer.
- The number of characters excluding URLs (Study 2 only).
- The number of URLs.[39]
- A measure of "uniqueness" of the approach.
  - Study 1: Three different research assistants (RAs) coded both individual and group answer sheets by the "creativity/uniqueness" of the steps used.[40] In order to make sure we are capturing answer sheets that stand out in their steps, we define "Different" as an answer sheet flagged as different by **at least two** RAs. Thus, in a committee of 3, the majority would consider the candidate answer sheet as "Different".[41]
  - Study 2: In order to capture the uniqueness of a solution approach we use URLs. For all individual answer sheets we count the number of occurrences of each URL. A URL that was unique gets assigned a 1, a URL that was used 40 times, a 40. As individuals may

[39] We excluded all URLs that refer to google searches. Often participants copied a link to the google search and also a link to an actual website. The google search represented an interim step and would thus double count sources.

[40] We asked the RAs to study a subset of answer sheets to learn about standard approaches taken for each guesstimation. Then they coded all answer sheets as "Different" that included steps that were sufficiently different from these standard approaches. The correlations between the coding of pairs of RAs are relatively low, but positive, on the order of $.3 - .4$.

[41] Over all treatments, 6% of group answer sheets and 9.8% of individual answer sheets are coded as Different. As an example, one answer sheet contained the step that there were more weddings due to the special date "06-06-06" in the year 2006 in June.

have copied multiple URLs, we average this uniqueness measure to create our Different measure per individual. We proceed analogously for group answer sheets.[42]

- A dummy whether the answer sheet is completed (disregarding the final answer itself), subjectively coded by an RA (Study 1 only).

- The identity of the individual who filled out the group answer sheet. In Study 1, we inferred this person by comparing the handwriting on the individual and the group answer sheets. In Study 2, individuals could take turns editing the group answer sheet. We thus count the number of times an individual edited the group answer sheet to construct the share of edits of each group member.

Work time

- Time of submission of each individual (Study 2 only) and group answer sheet.[43]

- Whether a timeout occurred (Study 2 only).

- An indicator of whether the group discussed unrelated matters during their work time (Study 1 only).

Group process measures

We collected further measures of group process.[44]

- Who presented their whole method (all steps on the individual answer sheet) and, if more than one participant did so, in which order.

- Communication measures:

[42]Manually we looked through URLs and adjusted them if they pointed to the same information. For example, URLs `https://www.transfermarkt.com/bundesliga/startseite/wettbewerb/L1/plus/?saison_id=` and `https://www.transfermarkt.de/bundesliga/startseite/wettbewerb/L1` were counted to be the same URL. This was often due to a .com and .de domain difference or extra characters at the end of the URL.

[43]In Study 1 this is the time from which on the group did not work on the guesstimation any more.

[44]In Study 1 we used video analysis. We had 1 RA code all 210 videos, which is the data we use here. Another RA coded a subsample of 129 videos which allowed us to test the reliability of the coding. Note that 21 out of 231 videos were lost because of technical problems. While RAs were not informed about the purpose of the experiment, and were thus blind to treatment, they did observe that in some videos participants filled out an additional sheet (the peer evaluation) at the end of the group phase.

  - Study 1: The number and order of individual speaking turns (only significant contributions count, no simple gesture of agreement, see Supplementary Appendix for details). From this we also calculate a participant's share of speaking turns.
  - Study 2: A count of each individual's messages posted in the chat and a count of words typed in the chat. From this we calculate the individual share of messages and words in the group chat.

Group atmosphere and motivation

We also elicited measures of participant's motivation and perceived group atmosphere through a survey at the end of the experiment:

- **Survey group atmosphere** The survey contains agree - disagree statements on a scale from 1 to 5 (fully disagree - fully agree) about the group environment. The following statements were used:
  - "I wanted to make a good impression on my group members."
  - "The atmosphere in the group was helpful."
  - "The atmosphere in the group was competitive."[45]
  - "I felt that everyone had an opportunity to voice their ideas in a fair way." (Study 1 only)
  - "I felt that others dominated the discussion in unproductive ways."
  - "All in all, my team did its best to come up with a good answer." (Study 2 only)
  - "I worked hard to come up with the best possible answer." (Study 2 only)
  - "During my work on this task, I felt stressed (in a negative way)." (Study 2 only)
- Preference for incentive scheme (Study 2 only): "How do you feel about the sharing rule? In this session it is determined by VOTE (PEER EVAL)/ RANDOMLY (BASELINE) which team member gets 50% / 30% / 20% of the team reward. If you could choose between a session

[45]In Study 1, we also stated: "Do you feel that competitiveness helped the group reach a better performance?" Participants that disagreed with the atmosphere being competitive did not always answer this question, and thus we exclude it from the analysis.

that uses this procedure and a session in which the 50% / 30% / 20% shares are allocated RANDOMLY (BASELINE) / by VOTE (PEER EVAL) to the three team members, which type of session would you prefer to participate in?"

- In the PEER EVAL treatment we asked whether an individual voted strategically or sincerely (unincentivized) and to give a reason for their vote (open text field).

Confidence measures

In Study 2 we elicited a participant's confidence in their preparation guess quality after they submitted their guess (incentivized).

- Absolute confidence: We ask them to predict the error band of their guess.
- Relative confidence: We ask them to predict the rank of their guess in their group.
- Vote confidence: We ask them to predict the number of votes they will receive. (PEER EVAL only)

A.2 Stereotype question

In Study 2, we asked participants after the experiment whether they thought that the main tasks (Labradors and Football) were easier for either men or women. They could indicate their views using a slider that, when shifted to the left, allowed them to select a score to indicate a male-favoring bias (100 being the most extreme possibility), and when shifted to the right, do the equivalent for women (again with 100 as the endpoint). For the Labrador task, most participants (59%) selected 0, the gender-neutral score, 26% thought that the task was easier for women, and 15% believed it was easier for men. Scores indicating a gender bias were on average of similar magnitudes in either direction (median value: 20 on the 0 to 100 scale on both sides). Overall, the null hypothesis of no difference is marginally rejected (p = 0.085). This, in turns out, is driven by female respondents: On average, they see a greater advantage for women, and when we consider only them, the null hypothesis is rejected with a p-value of 0.006 whereas male respondents see the task as gender-neutral (p = 0.727). In stark contrast to this, only 18% considered the Football task to be equally difficult for men or women, while 78% suggested that it favored men, and not much more than a handful, 4%, felt that women had an advantage. The median score on the male-favoring side is 51 and the null hypothesis of no gender difference is strongly rejected ($p < 0.001$ for both male and female respondents).

A.3 Additional Tables and Graphs

Table 9: Baseline characteristics by treatment and study

	Study 1		Study 2	
	BASELINE	PEER EVAL	BASELINE	PEER EVAL
Observations	93	78	99	102
Demographics				
Male	0.67(0.471)	0.56 (0.496)	0.55 (0.498)	0.55 (0.498)
Age	21.37(2.641)	21.06 (2.457)	24.40 (6.332)	23.25 (3.130)
Local Student	0.69(0.463)	0.78 (0.413)	0.22 (0.416)	0.19 (0.389)
Economics/Business Student	0.82(0.386)	0.87 (0.334)	0.64 (0.481)	0.59 (0.492)
STEM Student	0.06(0.246)	0.06 (0.245)	0.15 (0.359)	0.08 (0.269)
No Student	0.00(0.000)	0.00 (0.000)	0.04 (0.197)	0.04 (0.194)
Bachelor 1	0.25(0.431)	0.29 (0.456)	0.14 (0.348)	0.12 (0.322)
Bachelor 2	0.23(0.418)	0.23 (0.421)	0.11 (0.314)	0.08 (0.269)
Bachelor 3+	0.30(0.459)	0.22 (0.413)	0.20 (0.402)	0.24 (0.424)
Master or Higher	0.23(0.418)	0.26 (0.437)	0.51 (0.500)	0.53 (0.499)
Previous Experience Task	0.10(0.296)	0.13 (0.334)	0.25 (0.434)	0.28 (0.451)
Social Value Orientation				
Prosocial/Altruistic	0.61(0.488)	0.57 (0.496)	0.48 (0.500)	0.50 (0.500)
Big 5 Inventory				
Extraversion	6.79(1.757)	7.19 (1.721)	6.27 (1.765)	6.54 (2.151)
Agreeableness	7.29(1.441)	7.33 (1.345)	7.03 (1.475)	7.01 (1.551)
Conscientiousness	7.48(1.593)	7.29 (1.691)	7.14 (1.493)	7.69 (1.469)
Neuroticism	5.12(2.141)	4.87 (2.134)	6.23 (1.966)	5.99 (1.988)
Openess to Experience	6.88(1.621)	6.92 (1.673)	6.90 (1.604)	7.00 (1.723)
Ability Guesstimation				
Ability Guess Pay	3.20(2.995)	3.18 (3.230)	2.36 (3.373)	2.37 (3.568)

Note: The table reports group means. Standard deviations are reported in parentheses. One person in Study 2 stated non-binary as their gender. When considering heterogeneity by gender, we pool non-binary with female to avoid excluding the participant. Missing values are excluded from the group means for each characteristic separately. We have two missing values for each Big-5 item, both in Study 2 BASELINE, 1 missing value for Extraversion, Agreeableness and Conscientiousness in Study 1 BASELINE and three missing values of the SVO measure (all in Study 1, 1 in BASELINE, 2 in PEER EVAL). In the regressions we include these observations using mean imputation. Accounting for multiple hypothesis testing (separately by study), using the STATA rwolf2 command, we find no significant differences by treatment.

Table 10: Determinants of individual performance

	Ability stage		Preparation stages	
	Study 1	Study 2	Study 1	Study 2
PEER EVAL			−2.067 (1.334)	−0.529 (1.073)
Age	0.116 (0.152)	−0.074 (0.052)	−0.490* (0.261)	-0.015 (0.075)
Male	0.256 (0.561)	1.070** (0.524)	1.757 (1.292)	3.647*** (1.053)
Local	−0.017 (0.643)	0.577 (0.621)	1.438 (1.378)	0.050 (1.540)
Econ/Business	−0.380 (0.938)	0.924 (0.567)	−0.774 (1.653)	−1.599 (1.616)
STEM	−0.172 (1.359)	1.768** (0.882)	−0.557 (3.566)	1.986 (1.325)
Study Level	−0.222 (0.331)	−0.153 (0.210)	0.432 (0.785)	−0.399 (0.544)
High SVO Score	0.118 (0.504)	0.039 (0.488)	−0.561 (1.014)	−0.326 (0.911)
Experience	−0.223 (0.792)	0.809 (0.551)	−1.837 (1.150)	0.955 (1.402)
Extraversion	0.073 (0.150)	0.043 (0.130)	-0.080 (0.319)	0.0780 (0.266)
Agreeableness	0.288 (0.185)	−0.162 (0.163)	−0.229 (0.484)	0.050 (0.316)
Conscientiousness	0.121 (0.156)	0.119 (0.167)	−0.354 (0.305)	0.164 (0.318)
Neuroticsm	−0.106 (0.127)	−0.215 (0.135)	−0.284 (0.289)	−0.297 (0.330)
Openness	−0.005 (0.159)	−0.126 (0.151)	0.079 (0.323)	0.053 (0.436)
N	171	201	513	402
Clusters			11	20

Dependent variable: Achieved individual payoffs. The table reports OLS estimates with standard errors in parentheses (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). Columns 3 and 4 report robust standard errors clustered on the session level and corrected using wild bootstrap. Specification 4 (Study 2 preparation stages) contains an additional question/round indicator.

Table 11: Measures of effort during the preparation phase - pooled

Panel A: All

	(1) Timeout	(2) Steps	(3) Different
PEER EVAL	0.0982*	0.0595	0.00140
	(0.0507)	(0.129)	(0.0222)
Additional cov.	Yes	Yes	Yes
Observations	915	915	915

Panel B: Females

	(1) Timeout	(2) Steps	(3) Different
PEER EVAL	0.184***	0.155	0.0171
	(0.0520)	(0.230)	(0.0284)
Additional cov.	Yes	Yes	Yes
Observations	377	377	377

Panel C: Males

	(1) Timeout	(2) Steps	(3) Different
PEER EVAL	0.0584	-0.0347	-0.0155
	(0.0638)	(0.180)	(0.0323)
Additional cov.	Yes	Yes	Yes
Observations	538	538	538

The table reports OLS estimates for *Steps* and probit estimates (marginal effects) for *Timeout* and *Different*, with robust standard errors clustered on the session level for both studies. Asterisks (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$) refer to the treatment comparison (PEER EVAL vs. BASELINE). All specifications contain the full set of individual control variables shown in Table 9. We also include a question/round indicator. For *Different* in Study 2 created a dummy for answer sheets that have a uniqueness score in the top 25% of all preparation sheets containing a URL. For *Timeout* we used the variable *Timeout* in Study 2 and *Prep unfinished* in Study 1. For Panel A we conducted the STATA rwolf2 correction for multiple hypothesis testing. Our results are robust to this.

Table 12: Measures of effort during the group phase - pooled

Panel A: All

	(1) Work time	(2) Timeout	(3) Steps	(4) Turns	(5) Different	(6) Methods
PEER EVAL	0.940***	0.154**	0.0600	3.838**	0.0529	-0.0150
	(0.237)	(0.0607)	(0.234)	(1.779)	(0.0461)	(0.222)
Additional cov.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	296	296	305	296	305	296

Panel B: Majority-female teams

	(1) Work time	(2) Timeout	(3) Steps	(4) Turns	(5) Different	(6) Methods
PEER EVAL	0.454	0.0613	0.181	6.092**	0.209**	0.314
	(0.381)	(0.118)	(0.384)	(2.816)	(0.0955)	(0.375)
Additional cov.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	110	110	113	110	113	110

Panel C: Majority-male teams

	(1) Work time	(2) Timeout	(3) Steps	(4) Turns	(5) Different	(6) Methods
PEER EVAL	1.358***	0.216***	0.0202	3.237	-0.0127	-0.213
	(0.316)	(0.0473)	(0.310)	(2.135)	(0.0431)	(0.183)
Additional cov.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	186	186	192	186	192	186

The table reports OLS estimates for *Work time*, *Steps*, *Turns* (communication turns) and *Methods* (methods shared) and probit estimates (marginal effects) for *Timeout* and *Different*, with robust standard errors clustered on the session level for both studies. Asterisks (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$) refer to the treatment comparison (PEER EVAL vs. BASELINE). All specifications contain the full set of group control variables shown in the table notes of Table 1. We also include a question/round indicator. For *Different* in Study 2 created a dummy for answer sheets that have a uniqueness score in the top 25% of all group sheets containing a URL. For Panel A we conducted the STATA rwolf2 correction for multiple hypothesis testing. Our results are robust to this.

Table 13: Word groups and treatment (absolute)

	(1) Labradors Topic b/se	(2) Football Topic b/se	(3) Http b/se	(4) Positive b/se	(5) Negative b/se	(6) Leadership b/se	(7) Time Pressure b/se
PEER EVAL	2.529**	0.635	0.091	0.927*	0.040	0.509***	0.057
	(0.864)	(0.977)	(0.089)	(0.402)	(0.044)	(0.105)	(0.029)
Male	2.307**	2.699**	0.127	-0.073	0.061	-0.012	0.033
	(0.794)	(1.087)	(0.127)	(0.250)	(0.060)	(0.177)	(0.027)
Additional Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	201	201	402	402	402	402	402
Clusters	20	20	20	20	20	20	20

* p<0.10, ** p<0.05, *** p<0.01

Dependent variable: Number of words in a group member's chat messages belonging to a certain word group. The table reports OLS estimates with robust standard errors clustered on the session level and corrected using wild bootstrap (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). The full set of individual control variables shown in Table 9 and a question/round indicator. We also include indicators for the number of males in the group.

Table 14: Word groups and treatment (relative)

	(1) Labradors Topic b/se	(2) Football Topic b/se	(3) Http b/se	(4) Positive b/se	(5) Negative b/se	(6) Leadership b/se	(7) Time Pressure b/se
PEER EVAL	0.512	-0.944	0.011	0.616	0.008	0.354***	0.029
	(0.702)	(0.724)	(0.097)	(0.387)	(0.042)	(0.086)	(0.026)
Word count	0.063***	0.055***	0.003***	0.010**	0.001***	0.005***	0.001**
	(0.009)	(0.006)	(0.001)	(0.002)	(0.000)	(0.001)	(0.000)
Male	1.376*	0.629	0.064	-0.319	0.036	-0.135	0.011
	(0.626)	(0.810)	(0.132)	(0.236)	(0.059)	(0.182)	(0.026)
Additional Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	201	201	402	402	402	402	402
Clusters	20	20	20	20	20	20	20

* p<0.10, ** p<0.05, *** p<0.01

Dependent variable: Number of words in a group member's chat messages belonging to a certain word group. The table reports OLS estimates with robust standard errors clustered on the session level and corrected using wild bootstrap (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). The full set of individual control variables shown in Table 9 and a question/round indicator. We also include indicators for the number of males in the group.

Table 15: Preferences between BASELINE and PEER EVAL

	Prefer PEER EVAL	
PEER EVAL	1.464***	(0.362)
PEER EVAL * Male	0.432	(0.472)
Classified as prosocial	0.472**	(0.224)
Extraversion	0.0985	(0.0860)
Agreeableness	0.00605	(0.0693)
Conscientiousness	0.0762	(0.0986)
Neuroticism	-0.0111	(0.0775)
Openness to Experience	0.113	(0.0808)
Age	0.0481	(0.0322)
Male	-0.121	(0.401)
Dutch/German	-0.153	(0.270)
Study level	-0.124	(0.128)
Experience with guesstimations	-0.224	(0.305)
STEM	0.558	(0.624)
Economics or Business	0.529*	(0.290)
Ability payoff	0.0992**	(0.0431)
Observations	197	
Clusters	20	

* p<0.10, ** p<0.05, *** p<0.01

Note: The table reports ordered logistic regression estimates of the determinants of the preference for PEER EVAL. We lost 4 observations due to technical problems at the end of one session. Robust standard errors clustered by session are reported in parentheses. Asterisk indicate significance at the 10/5/1 percent level.

A.4 Group Chat Word Clouds

Word clouds are created from the corpus of all chat communication by first removing stop words using Python's nltk.corpus package (stopwords module) and then using the WordCloud package (Mueller, 2023).

Figure 4: Word cloud BASELINE

Figure 5: Word cloud PEER EVAL

References

Anderson, P. M. and Sherman, C. A. (2010). Applying the FERMI estimation technique to business problems. *The Journal of Applied Business and Economics*, 10(5):33–42.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, 76:451–469.

Ärlebäck, J. and Albarracín, L. (2019). The use and potential of Fermi problems in the STEM disciplines to support the development of twenty-first century competencies. *ZDM Mathematics Education*, 51:979–990.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, 14:1372–1400.

Bagues, M., Sylos-Labini, M., and Zinovyeva, N. (2017). Does the gender composition of scientific committees matter? *American Economic Review*, 107(4):1207–38.

Balietti, S., Goldstone, R. L., and Helbing, D. (2016). Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 113(30):8414–8419.

Bandiera, O., Fischer, G., Prat, A., and Ytsma, E. (2021). Do women respond less to performance pay? building evidence from multiple experiments. *American Economic Review: Insights*, 3(4):435–54.

Bohl, D. L. (1996). Minisurvey: 360-degree appraisals yield superior results, survey shows. *Compensation & Benefits Review*, 28(5):16–19.

Born, A., Ranehill, E., and Sandberg, A. (2022). Gender and willingness to lead: Does the gender composition of teams matter? *The Review of Economics and Statistics*, 104(2):259–275.

Boyle, I. (2013). Individual performance management: A review of current practices. *Asia Pacific Management and Business Application*, 1:157–170.

Bracha, A. and Fershtman, C. (2013). Competitive incentives: Working harder or working smarter? *Management Science*, 59(4):771–781.

Bracken, D. W. and Rose, D. S. (2011). When does 360-degree feedback create behavior change? and how would we know it when it does? *Journal of Business and Psychology*, 26:183–192.

Bradler, C., Neckermann, S., and Warnke, A. J. (2019). Incentivizing creativity: A large-scale experiment with performance bonuses and gifts. *Journal of Labor Economics*, 37(3):793–851.

Cahlíková, J., Cingl, L., and Levely, I. (2019). How stress affects performance and competitiveness across gender. *Management Science*, 66(8):3295–3310.

Cai, X., Lu, Y., Pan, J., and Zhong, S. (2019). Gender gap under pressure: Evidence from China's national college entrance examination. *The Review of Economics and Statistics*, 101(2):249–263.

Carpenter, J., Matthews, P. H., and Schirm, J. (2010). Tournaments and office politics: Evidence from a real effort experiment. *American Economic Review*, 100(1):504–17.

Chakraborty, P. and Serra, D. (2023). Gender and Leadership in Organisations: The Threat of Backlash. *The Economic Journal*.

Charness, G. and Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association*, 17:454–496.

Coates, D. E. (1998). Don't tie 360 feedback to pay. *Training*, 35(9):68–78.

Coffman, K., Flikkema, C. B., and Shurchkov, O. (2021). Gender stereotypes in deliberation and team decisions. *Games and Economic Behavior*, 129:329–349.

Cohen-Zada, D., Krumer, A., Rosenboim, M., and Shapir, O. M. (2017). Choking under pressure and gender: Evidence from professional tennis. *Journal of Economic Psychology*, 61:176 – 190.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668.

Delfgaauw, J., Dur, R., and Souverijn, M. (2020). Team incentives, task assignment, and performance: A field experiment. *The Leadership Quarterly*, 31(3):101241.

DeNisi, A. S. and Kluger, A. N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? *The Academy of Management Executive (1993-2005)*, 14(1):129–139.

Eckartz, K., Kirchkamp, O., and Schunk, D. (2012). How do incentives affect creativity? CESifo Working Paper Series No. 4049.

Edwards, M. R. and Ewen, A. J. (1996). How to manage performance and pay with 360-degree feedback: Multisource assessment can work for both performance and pay management when participants know the system is fair. But doing it right requires a commitment. *Compensation & Benefits Review*, 28(3):41–46.

Englmaier, F., Grimm, S., Schindler, D., and Schudy, S. (2017). The effect of incentives in non-routine analytical teams tasks - Evidence from a field experiment. CESifo Working Paper Series No. 6903.

Exley, C. L. and Kessler, J. B. (2022). The Gender Gap in Self-Promotion*. *The Quarterly Journal of Economics*, 137(3):1345–1381.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Frey, B. S. and Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5):589–611.

Gall, T., Hu, X., and Vlassopoulos, M. (2023). Incentivizing team leaders: A firm-level experiment on subjective performance evaluation of leadership skills. IZA Discussion Papers 16123.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, (1).

Harbring, C. and Irlenbusch, B. (2011). Sabotage in tournaments: Evidence from a laboratory experiment. *Management Science*, 57(4):611–627.

Hardt, D., Mayer, L., and Rincke, J. (2023). Who does the talking here? The impact of gender composition on team interactions. CESifo Working Paper No. 10550.

Huang, Y., Shum, M., Wu, X., and Xiao, J. Z. (2019). Discovery of bias and strategic behavior in crowdsourced performance assessment. https://arxiv.org/abs/1908.01718.

Isaksson, S. (2018). It takes two: Gender differences in group work. Working Paper.

Laske, K. and Schröder, M. (2017). Quantity, quality and originality: The effects of incentives on creativity. Technical Report 168151.

Lerchenmueller, M. J., Sorenson, O., and Jena, A. B. (2019). Gender differences in how scientists present the importance of their research: observational study. *Bmj*, (367).

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology*, 66:451–457.

Mueller, A. C. (2023). Wordcloud.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8):771–781.

Nalbantian, H. R. and Schotter, A. (1997). Productivity under group incentives: An experimental study. *The American Economic Review*, 87(3):314–341.

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1):601–630.

Ramm, J., Tjotta, S., and Torsvik, G. (2013). Incentives and creativity in groups. CESifo Working Paper Series No. 4374.

Sonnentag, S. (1998). Identifying high performers: Do peer nominations suffer from a likeability bias? *European Journal of Work and Organizational Psychology*, 7(4):501–515.

Teplitskiy, M., Ranu, H., Gray, G., Menietti, M., Guinan, E., and Lakhani, K. R. (2019). Do experts listen to other experts? Field experimental evidence from scientific peer review. Harvard Business School Working Paper, No. 19-107.

van Dijk, F., Sonnemans, J., and van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review*, 45(2):187 – 214.

Villeval, M. C. (2020). *Performance Feedback and Peer Effects*, pages 1–38. Springer International Publishing, Cham.

Weinstein, L. and Adam, J. A. (2008). *Guesstimation: Solving the World's Problems on the Back of a Cocktail Napkin*. Princeton University Press.

Zenger, T. R. and Marshall, C. R. (2000). Determinants of incentive intensity in group-based rewards. *The Academy of Management Journal*, 43(2):149–163.