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

John Morgan
University of California, Berkeley
Haas School of Business
morgan@haas.berkeley.edu

Susanne Neckermann
University of Chicago
sneckermann@uchicago.edu

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

This version: July 3, 2020.

Today's employees often work in teams on complex problems. Yet, we know very little about how to incentivize such work. Using a laboratory experiment, we study how groups performed on guesstimation problems under the following incentive schemes: fixed pay, pay that depends on group performance, and pay that depends on group performance as well as on the outcome of a peer evaluation. Overall, group performance was not affected by either kind of incentive. Yet, groups behaved signicantly differently under peer evaluation. Subjects reported higher motivation and groups worked harder and longer in the group phase. Our results suggest that subjects based their peer evaluation not on the actual correctness of an individual's answer but instead used more easily observable proxies of effort. Exploratory analyses suggest that female majority groups performed signicantly worse under peer evaluation, while the performance of male majority teams was not affected. We conclude that peer evaluations are an imperfect incentive mechanism when group members cannot easily assess the quality of individual performance.

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, Bart Francke, Jeroen Frijters, Colin Huliselan, Eline Korndewal, Jeannine van Reeken, Tessa Verloop and Zazie Weiffenbach for excellent research assistance. We also would like to thank Anne Boring, 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 U Lüneburg and the CBESS conference in East Anglia.

†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 whether and 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). 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 be used for developmental and confidential feedback only, many firms use this information for performance management. In fact, Edwards and Ewen (1996) report that more than 90% of those AMA member companies that use peer evaluations also use it 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).[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. (2017) 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 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

complex tasks, Ramm et al. (2013) and Englmaier et al. (2017). Neither considers peer evaluations.

To examine the effects of group and individual incentives (in the form of a peer evaluation) on group performance in a complex task, we performed a series of laboratory experiments where groups of subjects had to solve complex problems under varying incentive schemes. To capture problems relevant in business contexts, we chose guesstimations as our representation of complex problems. Guesstimation questions ask subjects to estimate some unknown, and highly unusual, quantity.[3] 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 and consequently many firms use these tasks in interviews of employees, seeing them as measuring critical problem solving abilities needed for their workforce (e.g. Anderson and Sherman, 2010).[4] They are also used in higher education. Solving guesstimations (also known as Fermi problems in the literature) 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 have definite, known, answers, so it is possible to grade performance in an objective fashion.

First, we examine the simple question of whether paying a group of three subjects on the basis of guesstimation accuracy improves the accuracy of the group's guesses relative to paying a fixed wage. Depending on the quality of their guess, a group could earn up to 35 euro under group incentives relative to a sure 15 euro under fixed pay.[5] Second, we study the effect of adding a peer evaluation to group incentive pay to determine the share of individual earnings. We do not only look at the effect on performance but also on what factors and behaviors result in positive performance evaluation, i.e. who is rewarded.

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. (2017), for example, study team incentives in a field experiment.

[3] A typical guesstimation would ask an individual to estimate the litres of toothpaste that were consumed in the United Kingdom in one year (our question 2).

[4] 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.

[5] We chose the fixed wage such that average pay was approximately equal to the other treatments.

The flow of the experiment was roughly as follows: We first asked each subject to solve a guesstimation question, which we use as an ability measure. Next, we grouped her with two other subjects, under a given incentive structure, and asked a new question. Within this group phase, individuals were first given 5 minutes to solve this question alone and then were given 10 minutes for a group discussion which resulted in the final group answer. Thus each group produced three individual answers (unincentivized) and one final group answer (incentivized depending on treatment). There was no structure on group discussion. We provided subjects with a sheet on which they could write down the steps used in deriving a given guesstimate. We then rematched subjects to new groups and repeated this group exercise twice more, each time with a new question. Throughout each session, we kept the incentive scheme constant and thus we study how performance varies with incentives between subjects.

Our main findings are:

1. Paying the group based on the accuracy of its estimates produced a small effect on group effort (measured by work time, number of speaking turns, conversation about unrelated matters and uniqueness of the answer steps), no effect on motivation, and no greater accuracy than paying a flat rate independent of accuracy.

2. Adding a peer evaluation to group incentives increased self-reported motivation to impress ones' peers and turned the perceived group atmosphere more competitive. It also increased group effort. Ultimately though, group performance was unaffected. Group guesses were as accurate under the peer evaluation as under group incentives only.

3. The failure of the peer evaluation to increase performance seems to stem from the fact that team members did not reward individual guess quality but rather more easily observable output measures such as the uniqueness and number of steps on the answer sheet or process measures such as the share of speaking turns.

Relating our findings to the literature on group complex problem solving, Ramm et al. (2013), studying a "eureka", thinking outside the box task (candle task), also find that incentives on the group level did not improve group performance. Englmaier et al. (2017), in contrast, find that

teams that had to escape an exit room did so more quickly when incentivised. In our study, subjects were not indifferent to incentive strength – they did increase their effort in the presence of a peer evaluation – this increase in effort, however, did not translate into an increase in performance. Our findings are consistent with the idea that peer evaluations encourage effort related to standing-out in the group with no additional benefit in terms of increased group performance. These varied findings suggest that this nascent literature needs more work to fully understand the link between performance and pay in group complex problem solving.

Our research also relates to the small empirical literature on peer evaluations or peer reviews in organizational research.[6] 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. Using an individual real effort task they show that workers expect to be underrated by their peers and therefore reduce their effort. As far as we are aware, we are the first to study the effects of pay tied to a peer evaluation in a complex group task using a laboratory experiment. In our set-up, the group is rewarded according to group output and the reward is split amongst team members according to the peer evaluation. Such a combination of group and individual incentives is typical in many workplace settings, where a team benefits from performing well, but each individual also benefits from being evaluated better than their peers.

Our findings also 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, which have previously been identified as impairing the effectiveness of peer evaluations (Edwards and Ewen, 1996, Coates, 1998), were minimized in our experimental setting, we identify a further limitation. When peers are not fully able to assess individual performance, behavior may shift towards more easily observable actions that are then rewarded by peers, though they do not necessarily improve performance. In our experiment subjects worked harder and longer, but at no

[6] In peer review the evaluator is typically not directly part of the work team producing the evaluated output, as for example in academic peer review for grant applications or journal submissions. Balietti et al. (2016) conduct a laboratory experiment to study in an art exhibition game how competition affects peer reviews as well as the innovativeness of the artworks. Teplitskiy et al. (2019) and Bagues et al. (2017) study how experts are influenced by other experts' evaluations in their (scientific) peer review. Huang et al. (2019) study actual performance evaluation data from a large service firm and document that raters underrate (overrate) competent (less competent) peers. Sol (2016), Deb et al. (2016) and Kim (2011) study the effectiveness of peer evaluations in a theoretical model.

additional benefit to the employer in terms of increased performance.

In light of persisting gender gaps on the labor market, gender differences in the response to incentives have been a topic of interest in the literature. Bandiera et al. (2017) show in a meta-study that males and females do not differ in their response to individual performance pay. Much less is known about group tasks and incentives relying on subjective performance evaluation by peers. Coffman et al. (2019) show that male and female contributions to teams are not equally rewarded in gender-stereotyped task. In an exploratory analysis on the role of gender, we compare the response of males and females to group and individual incentives. We find significant differences in the response to the introduction of a peer evaluation. While performance in male majority groups was largely unaffected, female majority groups performed significantly worse relative to a setting with group incentives only. We show that our results are broadly consistent with females choking under pressure and underperforming when facing evaluation by their team members.

Next, we describe the experimental design and the rationale for key design decisions. Section 3 summarizes our main results. In Section 4, we study gender differences in the response to incentives. Finally, Section 5 concludes.

2 Experimental Design

2.1 Experimental Design

We conducted 15 sessions at the econlab of the Erasmus School of Economics at Erasmus University in Rotterdam. Sessions occurred over the course of May and June 2014, each lasting about 120 minutes. Using ORSEE recruiting software, we obtained a total of 231 participants, consisting of a mix between graduate (25%) and undergraduate students (75%). No subject participated more than once.

We began each session by reading the instructions aloud to subjects who read along silently. We then read aloud a worked-out example of a guesstimation question and taught subjects 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 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). To aid their calculations as well as level the playing field, we provided subjects with a factsheet that included population numbers as well as with calculators. Subjects were informed that they would be videotaped.

Each experimental session consisted of 4 stages. In stage 1 each subject worked on a guesstimation question by themselves. Stages 2-4 were group stages where we implemented different incentive schemes. In each session we kept the incentive scheme constant throughout stages 2-4 and thus we identify treatment differences in a between-subjects design.[7] In the following, we present more details first on stage 1 and then on stages 2-4.

In stage 1 subjects were presented with an incentivized guesstimation question to work out themselves. This question was the same for all subjects. We used their performance in this task as a proxy for individual ability at solving guesstimations. A subject's earnings in this stage depended on the percentage error of their guess compared to the correct answer. We divide these percentage errors into six bands, each with a different payoff. If the answer is within 10% of the correct answer, they received the full prize. If they were off by plus/minus 20%, 40%, 60%, or 80% they received 80%, 60%, 40% or 20% of the prize. See Table 1 for an overview. In the individual task described above, the prize was 10 euro. For example, an individual that submitted a guess that was within 40% of the true answer received $0.6 * 10 = 6$ euro. Subjects received no feedback about the quality of their choice at any point during the experiment.

Following the individual guesstimation, subjects were randomly assigned into groups of three.[8] Each group was sent to a separate room to work on a guesstimation question. This stage 2 guesstimation question was the same for all groups. Groups had 15 minutes at their disposal to come up with a final answer. For the first five minutes, each member of the group worked alone at a separate desk facing the wall. We will later refer to this as the *individual phase*. Subjects were encouraged to write down an answer to the question and indicate the steps by which they arrived at this answer

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

[8] More precisely, each subject was privately and randomly assigned an identifier, which fully determined the room and table that this subject would occupy in each round. Subjects could find this information in the instructions available to them.

Table 1: Share of prize awarded depending on guess quality

Prize share	Construction
0	Guesstimation is more than +/- 80% of the true answer
0.2	Guesstimation is within +/- 80% of the true answer
0.4	Guesstimation is within +/- 60% of the true answer
0.6	Guesstimation is within +/- 40% of the true answer
0.8	Guesstimation is within +/- 20% of the true answer
1	Guesstimation is within +/- 10% of the true answer

Note: The construction of the incentive scheme used to reward the first individual guess as well as group guesses (depending on treatment) as used during the experiment and explained to the subjects.

by filling out a worksheet[9]. They were made aware that the completion of this worksheet was not incentivized. For the remaining 10 minutes, subjects worked collectively as a group at a large table in the middle of the room. We refer to this as the *group phase*. Groups were free to choose whatever process they wished to arrive at a solution. For instance, a group could choose a "chairman" to moderate the conversation. They could speak in turns or all at once, and so on. As in the individual phase, the group was provided with a worksheet where they needed to indicate their final answer. There was also room to indicate the steps taken in deriving this final answer. Rewards were only tied to the final answer.

Our treatments vary the payment structure to allow us to learn more about how performance in complex tasks varies by incentive scheme. In the FLAT treatment, subjects were paid a flat compensation of 15 euro per group (and thus 5 euro per individual) irrespective of the quality of the group guess, as long as they submitted an answer.[10] They received this compensation on top of their earnings from the individual guesstimation and other earning opportunities at the end of the experiment. In the GROUP INC treatment, group earnings in that round depended on the quality of the group guess and were calculated using the incentive scheme in Table 1 on a maximal prize of 35 euro. Thus, a group that submitted a guess that was within 40% of the true answer, for example,

[9] An example of a filled-in worksheet is provided in the Supplementary Appendix.

[10] The instructions stated: "For every round you will receive 5 Euros for submitting a group answer on the yellow sheet. This payment is independent of the quality of your answer. If your group does not submit an estimate, you do not earn anything in this round." Using data from sessions on the other two treatments, we chose the flat compensation such that average payment for the group task would be similar.

received $0.6 * 35 = 21$ euro. The group reward was then split in three shares of 50%, 30%, and 20%, respectively. At the end of the experiment, we randomly determined the allocation of these reward shares across the three group members. In the PEER EVAL treatment, we implemented a peer evaluation to split the group surplus. After the group submitted its guess, each member of the group privately voted for one of the other two members who they thought "contributed more to the solution to the problem". This vote was based on perceived performance, as subjects were not informed about the quality of their final guess. The members of the group received no immediate feedback as to the results of the vote. The group earnings from each round were calculated using the incentive scheme in Table 1 but now using a maximal prize of 35 euro, just as in GROUP INC. Of the total group earnings from this guesstimation the individual voted most valuable by both team mates received 50%, 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.[11] Note that we chose to implement noisy group incentives in GROUP INC to ensure that any differences between the PEER EVAL and GROUP INC treatment were driven by the peer evaluation and not by the earnings structure.

After solving a guesstimation question, but without receiving feedback on their performance, nor on the results of the peer evaluation in case of the PEER EVAL treatment, the subjects in a session were randomly reassigned into new groups of three, sent off to a new room, and given a new guesstimation question to solve. Again, this guesstimation question was the same for all subjects. The randomization scheme ensured that group members never overlapped across guesstimations (perfect stranger matching). Each session consisted of three group guesstimation questions, all under the same incentive scheme.[12]

Following the guesstimation phase of the experiment, subjects privately completed a question-

[11] 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 subject 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.

[12] Note that our matching procedure did not allow us to also randomize the order of the questions across rounds in a session. Thus we are not able to study learning effects over rounds. Since subjects received no feedback regarding their performance and never interacted with the same individuals twice, we expect these learning effects to be relatively small in our setting.

naire, parts of which were compensated, designed to collect information about group atmosphere, social value orientation (based on the slider measure of Murphy et al., 2011), personality traits (Big 5 index (BFI-10)), and demographics. Social value orientation, personality and demographics data, together with the proxy for individual skill at the task, were used as controls for estimating treatment effects.[13]

At the conclusion of each session, subjects were paid in private and in cash based on the quality of their guesses as well as their choices in the incentivized parts of the questionnaire. On average, subjects earned 21.15 euro for their participation, a rate meeting or exceeding the typical earning opportunities available to them.

The guesstimation questions we used were

- **Individual "Ability":** 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? (A: 2 million km)

The Supplementary Appendix contains reproductions of all of the materials presented to subjects, including instructions, the fact sheet (including the population sizes of the relevant countries), a sample answer sheet and the questionnaires.

2.2 Description of Variables

2.2.1 Performance Measures

We use two measures of performance:

[13]Note that answers to the survey questions may have been affected by performance earlier in the experiment and thus they may be endogenous to treatment. To minimize such distortions subjects received feedback about their performance only at the end of the experiment after they had filled in the survey.

- (Hypothetical)[14] group payoff (0-35 euro) calculated using the incentive scheme in Table 1 and the answer to the guesstimation.

- Percentage error (smaller numbers mean better performance):

$$\text{P.E.} = \frac{|\text{Guess} - \text{Truth}|}{\text{Truth}}$$

2.2.2 Other Outcome Variables

Aside from the final guesstimation answer, we collect 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.

- In order to capture the creativity of the steps used, we asked three different research assistants (RAs) to code both individual and group answer sheets by the "creativity/uniqueness" of the steps used.[15] 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".[16]

- The identity of the individual who filled out the group answer sheet (inferred by comparing the handwriting on the individual and the group answer sheets) as a proxy for leadership.

We collect further measures of group effort and process through a video analysis.[17]

[14] We refer to hypothetical group payoffs, because under the FLAT treatment groups were not rewarded by the quality of their answer. For this treatment, the calculated rewards are thus hypothetical and not actual.

[15] 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 $.3 - .4$.

[16] 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.

[17] We had 1 RA code all 210 videos, which is the data we use here. Another RA coded a subsample of 129 videos so that we could test the reliability of the coding. Correlations are between .47 (speaking turns) and .9 (work time). 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 subjects filled out an additional sheet (the peer evaluation) at the end of the group phase.

- Who presented their whole method (all steps on the individual answer sheet) and, if more than one subject did so, in which order.
- Who shared their individual guess and in which order did subjects present their individual guesses if more than one subject did so.
- The total work time of the group, calculated as the start time until stopping time from which moment on the group did not work on the guesstimation any more. Thus conversations about unrelated matters *during* the discussion are included in the work time.
- An indicator of whether the group discussed unrelated matters.
- 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 subject's share of speaking turns.
- A subjective assessment by the RA of whether all group members agreed with the final group answer.
- A subjective assessment by the RA of the dominance structure in the group (one, two or no dominant individual(s) and their identity).

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

- **Survey group atmosphere** The survey contains agree - disagree questions on a scale from 1 to 5 (fully agree - fully disagree) about the group environment. The following questions were asked:
  - 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
  - Do you feel that competitiveness helped the group reach a better performance?

– I felt that everyone had an opportunity to voice their ideas in a fair way

– I felt that others dominated the discussion in unproductive ways

Furthermore, in the PEER EVAL treatment we asked whether an individual voted strategically or sincerely (unincentivized).

3 Results

3.1 Sample and Summary Statistics

Table 2 summarizes the characteristics of our subjects by treatment.[18] There are no significant differences in subject characteristics between treatments, except for one variable. Subjects in GROUP INC and PEER EVAL exhibited a higher score for the openness to experience Big 5 personality measure (significant at the 10% level). Given that we are performing multiple tests, even in the absence of any real difference some significant coefficients are to be expected.[19] Nonetheless, in addition to analyzing average treatment differences we include the variables in Table 2 as control variables in all further individual-level regressions where indicated as the standard set of individual controls in order to reduce the variance of the estimated treatment effect.

Table 3 displays summary statistics of the individual and group guesses by guesstimation question (Cycling, Toothpaste and Weddings). The variance in guess sizes was large, warranting our choice of these guesstimations as complex problems. While the mean answer was much larger than the true answer in all three question, the median guess was relatively closer (i.e. a percentage error of 0.047, 0.377, 0.414 for Cycling, Toothpaste and Weddings respectively). We use this distribution to winsorize the guesses for further analysis at the 90% level for the *Percentage Error* performance measure to account for outliers.[20]

[18] Note that the number of observations differs between treatments. Due to administrative reasons with the subject-recruiting process we were not able to obtain a fully balanced sample.

[19] More precisely, we conduct a total of 36 tests and thus we should expect 3.6 tests to be significant at the 10% level even in the absence of real differences.

[20] Winsorisation is done by using the distribution of all guesses made for each question. This includes the group guesses as well as all individual guesses. A 90% Winsorisation is used, this means that all the guesses below the 5th percentile are set to the 5th percentile, and all guesses above the 95th percentile are set to the 95th percentile.

Table 2: Baseline Characteristics by Treatment

	FLAT	GROUP INC	PEER EVAL
Observations	60	93	78
Demographics			
Female	0.367(0.482)	0.337 (0.473)	0.436 (0.496)
Age	21.333(2.370)	21.391 (2.643)	21.064 (2.457)
Dutch	0.700(0.458)	0.685 (0.465)	0.782 (0.413)
Economics Student	0.767(0.423)	0.685 (0.465)	0.731 (0.444)
Econometrics Student	0.117(0.321)	0.065 (0.247)	0.064 (0.245)
Bachelor 1	0.283(0.451)	0.239 (0.427)	0.295 (0.456)
Bachelor 2	0.150(0.357)	0.228 (0.420)	0.231 (0.421)
Bachelor 3	0.333(0.471)	0.304 (0.460)	0.218 (0.413)
Master or Higher	0.233(0.423)	0.228 (0.420)	0.256 (0.437)
Previous Experience Task	0.117(0.321)	0.097 (0.296)	0.128 (0.334)
Average Grade	7.142(0.725)	7.182 (0.762)	7.096 (0.780)
Social Value Orientation			
	0.458(0.498)	0.391 (0.488)	0.434 (0.496)
Big 5 Inventory			
Extraversion	6.983(1.396)	6.793 (1.757)	7.192 (1.721)
Agreeableness	7.200(1.527)	7.293 (1.441)	7.333 (1.345)
Conscientiousness	7.169(1.544)	7.478 (1.593)	7.295 (1.691)
Neuroticism	4.700(2.011)	5.118 (2.141)	4.872 (2.134)
Openness to Experience	6.417(1.544)	6.882* (1.621)	6.923* (1.673)
Ability (Dog Question)			
	0.347(0.329)	0.324 (0.299)	0.318 (0.323)

Note: The table reports group means. Standard deviations are reported in parentheses. One person recorded bachelor 4 as year of study in the PEER EVAL treatment, this observation is pooled with bachelor 3. Four pre-master students (1 FLAT, 1 GROUP INC, and 2 PEER EVAL) and one PhD student (FLAT) are pooled with master students. Missing values are excluded from the group means for each characteristic separately. We have one missing value for gender, age, nationality, field of study, year of study, extraversion, agreeableness, conscientiousness, and ability in the FLAT group. Besides that, four missing values of the SVO measure (1 FLAT, 1 GROUP INC, 2 PEER EVAL), three missing values of average grade (all in FLAT). In the individual-level regressions we include these observations using mean imputation and including a missing indicator. We also elicited the number of quantitative classes but many subjects failed to report an answer leading to 17 missing values (10 FLAT, 7 PEER EVAL). Thus we exclude this variable. Asterisks indicate a difference of means (compared to FLAT) significant at the 10/5/1 percent level using a two-sided t-test. There are no significant differences between GROUP INC and PEER EVAL.

Table 3: Summary Statistics of the Guesstimations by Question

	Cycling (in 10,000)	Toothpaste (in 1,000,000)	Weddings (in 1,000)
# Observations	274	267	267
Mean	652.82	1,147.74	886.12
Maximum	39,647.06	242,027.00	52,000.00
Minimum	3.75	0.00	0.05
Standard deviation	2,671.40	14,868.54	4,184.87
1st Percentile	6.00	0.00	1.10
5th Percentile	37.64	0.33	4.00
10th Percentile	54.00	6.21	12.50
25th Percentile	109.55	30.00	33.44
50th Percentile	190.56	63.74	70.00
75th Percentile	421.20	127.49	295.64
90th Percentile	1,008.00	340.15	1,181.20
95th Percentile	1,900.00	570.02	2,551.40
99th Percentile	9,175.52	5,344.09	20,230.00
True Answer	200.00	46.30	49.50

Note: The sample of the table includes all group and individual guesses made for each of the questions (before winsorisation). This means, it includes 77 group guesses and the rest are individual guesses. Since not all individuals always made an individual guess, the number of observations are lower than 308 per question.

3.2 Do Incentives Matter?

3.2.1 Performance

Figure 1 plots the distribution of (hypothetical) reward shares that follow from applying the group incentive scheme to the group answer by treatment. These reward shares ranged from 0 to 1 in steps of .2, as explained in Table 1. No large differences in distributions are visible across treatments. The mean (hypothetical) payoffs (reward share times maximal prize of 35 euro) were 12.13 euro (FLAT), 12.95 euro (GROUP INC) and 13.2 euro (PEER EVAL), out of a possible 35 euro. Economically these differences are small, the maximal difference between FLAT and PEER EVAL equals 3% of the total earnings potential. Also the variance in payoffs is similar. For example, the share of groups earning the maximal prize equals 7%, 11% and 8% in FLAT, GROUP INC and PEER EVAL respectively, while the share of groups earning nothing equals 37% (FLAT), 33% (GROUP INC) and 33% (PEER EVAL).

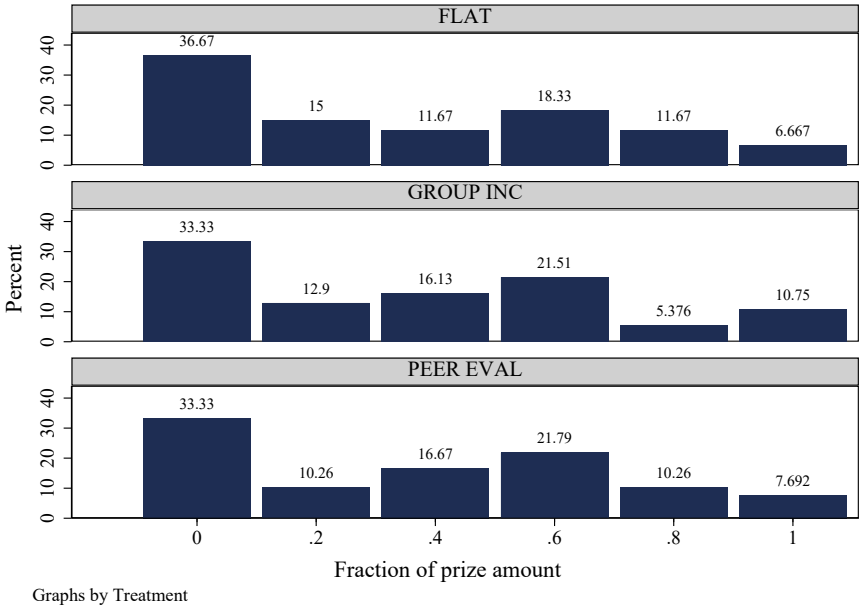

Figure 1: Distribution of the (hypothetical) share of rewards earned by groups in each treatment.

Table 4 reports the results of a regression analysis using both hypothetical group payoffs (between 0 and 35 euro) and winsorized percentage error as performance measures with and without control variables. Row "Group Incentives" shows the effect of going from FLAT to GROUP INC, row "Group Inc. & Peer Evaluation" the effect of going from GROUP INC to PEER EVAL and row "Peer Evaluation" the effect of going from FLAT to PEER EVAL (the sum of the previous two coefficients). Note that going from column (1) to column (2) by adding control variables slightly decreases the maximal effect size between FLAT and PEER EVAL from 1.1 to .8 euro (2.3% of the total earnings potential). Neither the effect of group incentives only, nor group incentives paired with a peer evaluation are statistically significant for either measure of group performance. Thus, we do not find evidence that monetary incentives on either the group or the individual level were able to foster performance. On the flipside, we also do not find any evidence of a dark side of incentives.

Why did incentives not affect performance? Ariely et al. (2009) argue that the effect of incentives on performance can be separated into two parts: a) the effect of increasing performance-contingent incentives on motivation and effort and b) the effect of greater motivation and effort on performance.

Table 4: Effects of Treatments on Group Performance

	Payoff		Percentage Error	
	(1)	(2)	(3)	(4)
Group Incentives	0.813	0.513	0.513	0.674
	(2.397)	(2.246)	(0.564)	(0.545)
Group Inc. & Peer Evaluation	0.246	0.286	-0.082	-0.391
	(1.402)	(2.030)	(0.784)	(0.583)
Additional Covariates		Yes		Yes
Peer Evaluation	1.059	0.799	0.431	0.283
Observations	231	231	231	231
Clusters	15	15	15	15

Note: The table reports OLS estimates of the impact of group and individual incentives on group payoff and percentage error. The percentage error columns use winsorized errors. Robust standard errors clustered on the session level are reported in parentheses. Additional covariates consist of the mean and maximum group ability score, a measure for group SVO, measures for group Big5 items, the number of females, the average age of group members, an indicator for groups with same nationality group members, percentage first and second year bachelor students, number of group members with experience with guesstimations, the number of Economics students, the number of Econometrics students and average as well as maximum grade point average. Asterisks indicate significance at the 10/5/1 percent level. P-values are calculated using wild bootstrap (STATA boottest) to account for the small number of clusters.

In our setting, we can further distinguish between effort in the individual (preparation) phase and effort in the group phase. In the following, we study the link between incentives and effort. We do so by studying a) the individual answer sheets in the individual phase of the group task, b) data from the video analysis of the group phase, and c) answers on motivation and atmosphere in the group from the questionnaire. Next, we address the link between effort and performance by studying voting behavior in the peer evaluation.

3.2.2 Effort

We begin by studying effort in the individual phase of the group task. During the first 5 minutes of the total of 15 minutes allocated to the group task, subjects had time to prepare by filling out the individual answer sheet. Table 5 presents measures of the fraction of individual answer sheets that contained an individual guess, the payoff this guess would have received under the group incentive scheme (only answer sheets that contain a final guess), the number of steps taken for all answer sheets (all), for answer sheets that contain a final guess (complete), for answer sheets that do not contain a final guess (missing), as well as whether the answer sheet is coded as "Different" (all answer sheets). The only significant difference between treatments is the fraction of missing individual guesses. This fraction increased from 8% to 24% when increasing incentives from FLAT to PEER EVAL. Thus, roughly a quarter of individual answer sheets did not contain a final guess under PEER EVAL.

Whether individual effort indeed increased with incentives is not clear. The increased fraction of missing guesses could have been due to less but could also be driven by more effort.[21] Arguably, if incentives discouraged effort, we might expect to see also a decrease in payoff, a decrease in steps and a decrease in the fraction of "Different" answer sheets. Table 5 shows no evidence of this. If incentives on the other hand encouraged effort, this could have led to individuals failing to finish on time. Whether this would be accompanied by an increase in payoff of answer sheets that are filled in, the number of steps or the number of "Different" answer sheets would depend on who these

[21] Alternatively, there could be strategic reasons for not writing down a final answer. We do not think this is likely in our setting. Subjects did not have to share their individual answer sheet with their group members and remembering a final answer without noting it down is difficult. Additionally, an RA subjectively coded whether the steps on an answer sheet seem completed (disregarding the final answer itself). The correlation between a missing guess and an answer sheet being coded unfinished is .88.

Table 5: Treatment Differences Individual Phase of Group Task

	FLAT	GROUP INC	PEER EVAL
Observations	180	279	234
Missing Guesses	0.08	0.17*	0.24**
Individual Payoff (complete)	8.39	10.26	8.56
Ind. Steps (all)	4.96	4.83	4.80
Ind. Steps (complete)	5.02	4.97	4.85
Ind. Steps (missing)	4.29	4.11	4.65
Different	0.09	0.10	0.11

Note: The table reports group means. Asterisks directly after a statistic indicate a difference of means compared to FLAT significant at the 10/5/1 percent level. No significant differences between GROUP INC and PEER EVAL were found. The p-values for payoff and steps are calculated from a regression clustering standard errors at the session level corrected using wild bootstrap (STATA boottest) to account for the small number of clusters. The p-values for different and missing guess are calculated from a probit clustering standard errors at the session level corrected using wild bootstrap (STATA boottest). The standard set of individual control variables is included in all specifications.

individuals were that did not complete the answer sheet. For example, if it was especially the most talented individuals that did not finish in time and thus did not write down an individual answer, we might see a reduction in payoff of completed answer sheets even if the quality of the answers of the less talented individuals did increase. Without digging deeper into this issue, a selection problem plagues the analysis. Given the individual phase measures it is thus hard to ascertain whether effort increased through incentives at this point.

To further examine this question, we analyze differences in behavior and effort in the group phase. Table 6 shows different group effort and process measures by treatment. Using a video analysis, we measured how many individuals shared their individual guess with the group, how many presented their whole methodology (i.e. all the steps on their answer sheet), the time until the group stopped working on the guesstimation, whether the group had a (significant) conversation about unrelated matters, the number of (significant) speaking turns in the group discussion, whether there was a dominant individual/two dominant individuals, whether it seemed that the group in the end agreed on the group answer and whether the group answer sheet was coded as "Different".

While the increase in missing guesses observed in Table 5 translated into less individual guesses being shared in GROUP INC (relative to FLAT), this turns out not to be statistically significant

Table 6: Treatment Differences Group Phase

	FLAT	GROUP INC	PEER EVAL	
Observations	60	93	78	
Guesses Shared	1.85	1.34**	1.49	
Methods Shared	1.10	1.20	1.08	
Worktime	7.97	8.37**	8.83**	*
Unrelated	0.69	0.49***	0.49	
Number of Speaking Turns	9.71	10.45	11.89*	
All Agree	0.88	0.86	0.83	
No Dominant Indiv.	0.38	0.33	0.37	
Different	0.05	0.04	0.09**	**

Note: The table reports group means. Asterisks directly after a statistic indicate a difference of means compared to FLAT significant at the 10/5/1 percent level. The last column compares GROUP INC and PEER EVAL. The p-values for guess, method, worktime and turns are calculated from a regression using robust standard errors clustered at the session level corrected using wild bootstrap (STATA boottest). The p-values for unrelated, all agree, dominant and different are calculated from a probit regression using robust standard errors clustered at the session level corrected using wild bootstrap (STATA boottest). The standard set of group control variables is included in all specifications.

for PEER EVAL (which featured even more missing individual guesses). The number of times a complete individual method (i.e. the sum of all steps to arrive at a solution) was presented, does not differ significantly by treatment. On average, just one subject presented their whole answer sheet in all treatments.[22] The work time of groups under the PEER EVAL treatment was on average nearly one minute longer than under FLAT, with GROUP INC in between. Groups had 10 minutes, so in general, most of the time was spent on task. Also, the chance of engaging in conversation on unrelated topics during work time was 20 percentage points lower under PEER EVAL (and GROUP INC) than under FLAT, though only the latter reaches statistical significance. The number of speaking turns taken was on average increased by 2 turns under PEER EVAL. We see no significant treatment difference in the chance of agreement at the end of the group phase nor in the occurrence of dominant individuals. Finally, the percentage of "Different" answer sheets nearly doubled under PEER EVAL relative to both GROUP INC and FLAT. Overall the results seems to suggest that groups worked harder in the group phase under PEER EVAL and also, though

[22]There was a lot of heterogeneity in how groups shared information about their individual answer sheets. Most of the groups did not share a complete method (42%) while a sizeable fraction of groups (19%) shared the method of all three group members.

somewhat less so, under GROUP INC.

How can we interpret this increase in effort? First, note that none of the group effort measures in Table 6 is an unambiguous measure of "productive" effort. While, for example, working longer on the task may be an indication of a larger investment of "productive" effort to make the group guess better, it may equally well be an indication of an increase in effort directed at standing out by each single group member. While trying to impress the group may indirectly be productive for the group, it does not necessarily have to be the case. It may even deteriorate information exchange as an individual trying to impress may be less open to listen to others' point of view or may be exaggerating the confidence in their own answer, thus distorting group decisions. Similarly, a "Different" answer sheet may help make the answer to the Guesstimation more accurate, but it may also serve to impress the group members with one's creativity. In the next section we study subjects' reported motivation and group perception to shed further light on this question.

3.2.3 Motivation

Table 7 reports the answers of subjects regarding their motivation and perception of the group atmosphere by treatment. While there are no significant differences between GROUP INC and FLAT, PEER EVAL seems to have significantly affected subjects' perceptions of the group phase. Subjects agreed significantly more with the statement "I wanted to make a good impression on my group members" in the PEER EVAL treatment relative to both other treatments. But not only their own motivation seemed to change, in general subjects perceived the atmosphere as more competitive. While this may or may not encourage a better performance, subjects also agreed more with the statement that "I felt that others dominated the discussion in unproductive ways" although the average subject still somewhat disagreed with this statement. This points towards a potential performance-reducing effect of the peer evaluation. To sum up, our findings are consistent with PEER EVAL motivating higher, though not necessarily more productive effort relative to both FLAT and GROUP INC.

Table 7: Perception of Group Environment by Treatment

	FLAT	GROUP INC	PEER EVAL	
I wanted to make a good impression on my group members	3.506	3.638	4.124***	***
The atmosphere in the group was helpful	4.161	4.090	4.158	
The atmosphere in the group was competitive	2.367	2.290	2.714*	***
Do you feel that competitiveness helped the group reach a better performance	3.123	3.281	3.142	
I felt that everyone had an opportunity to voice their ideas in a fair way	4.472	4.430	4.376	
I felt that others dominated the discussion in unproductive ways	1.911	1.792	2.047*	***

Note: Answers to the survey questions range from 1 fully agree - 5 fully disagree. The p-value are obtained from an ordered logit regression clustering standard errors at the session level corrected using wild bootstrap (STATA boottest). The standard set of individual control variables is included in all specifications. Significance stars are shown as follows: * p< 0.1, ** p< 0.05, *** p< 0.01. Asterisks directly after a statistic indicate a difference of means compared to FLAT significant at the 10/5/1 percent level. The last column compares GROUP INC and PEER EVAL. Note that the number of observations are lower for the 4th question because a percentage of the subjects who didn't perceive the group as competitive skipped this question.

3.2.4 Determinants of the Peer Evaluation

As a final step, we study individual voting behavior in the PEER EVAL peer evaluation. If good individual performance in the form of a precise individual answer was not rewarded, and this was anticipated by subjects, incentives for good performance on that dimension were absent. At the same time, if other sorts of behavior did lead to a higher chance of being voted the most valuable group member, behavior may have shifted in response to treatment without affecting performance.

We find that the "best" individual (i.e. the individual with the most precise individual answer) was only voted winner in the peer evaluation in 21.5% of cases. Note, that in 15.4% of cases, a tie was the group outcome (and thus no individual became winner of the peer evaluation). This suggests that other factors must have determined voting. We now study the determinants of the peer evaluation more carefully using regression analyses. Table 8 presents the results of a probit 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 (3) present the coefficients of probit regressions where each variable of interest is included *separately*. This is done as some variables measure similar concepts

and are thus moderately correlated.[23] Columns (2) and (4) present results where all variables of interest are included *simultaneously*. Columns (1) and (2) show results pertaining to the full sample, while columns (3) and (4) restrict the analysis to groups where no subject stated that they voted strategically (self-reported in the final questionnaire, unincentivized).[24] While the full sample is more informative of the actual incentives the peer evaluation induces, it is interesting to see whether individuals that reported to have voted non-strategically rewarded the best individual guess.

Several things are interesting to note. First, having had the best answer in the group is not significantly related to being voted most valuable team member. Good performance, measured by the quality of the individual guess, was not rewarded, not even amongst the subset of individuals that stated to vote non-strategically. Furthermore, not having an individual answer at all is also not related to the chance of being voted most valuable team member. On the other hand, other measures of individual effort that are easily observable are significantly and positively related to winning the peer evaluation at least in some specifications: the number of steps on the individual answer sheet, having a "Different" guess, being perceived as dominant, the share of speaking turns and whether an individual presented their method. These results, if anything, are slightly stronger in the sub-sample of individuals that reported to vote non-strategically.

Since no feedback was given regarding the quality of the group answer before the vote, our results are consistent with subjects considering the individual solution steps of each group member and how they were presented as a proxy for their group contribution in their vote.[25] Different to the number of steps, being dominant or the share of speaking turns, which are easily observed, a group member with a "Different" individual answer sheet needed to communicate the details of his or her steps to the group. If subjects anticipated being rewarded for their creativity, we would expect

[23] "Turns Share" and "Dominant" are the most highly correlated with a correlation coefficient of .41.

[24] The scope for strategic behavior was limited. The only effect of distorting one's vote is to turn a sure second place into a tie and vice versa. Even under full information about others' voting behavior the expected difference in payoffs between sincere and strategic voting is negligible.

[25] 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 quality of the final guess, the quality and creativity of the answer steps seem to be regarded as an important indication of this.

Table 8: Determinants of the PEER EVAL Vote

	Probit (ME) winner PEER EVAL			
	Full sample		Non-strategic	
	Separate	All	Separate	All
Best Guess	-0.14	-0.15	-0.06	-0.09
	(0.089)	(0.065)	(0.105)	(0.073)
Missing Guess	0.06	0.04	0.04	0.06
	(0.066)	(0.030)	(0.066)	(0.023)
Ind. Steps (all)	0.05*	0.05*	0.07***	0.06
	(0.019)	(0.023)	(0.021)	(0.027)
Different Guess	0.14***	0.12*	0.22***	0.20*
	(0.037)	(0.045)	(0.054)	(0.060)
Filled Out Group Answer Sheet	-0.01	-0.03	0.05	0.03
	(0.058)	(0.060)	(0.076)	(0.083)
Turns Share	0.58*	0.31*	0.83***	0.39
	(0.225)	(0.174)	(0.276)	(0.277)
Dominant	0.13	0.07	0.16	0.11***
	(0.060)	(0.054)	(0.085)	(0.038)
Presented Guess	0.07	0.04	0.16	0.10
	(0.056)	(0.061)	(0.085)	(0.091)
Presented Method	0.05	0.01	0.15*	0.10
	(0.031)	(0.040)	(0.043)	(0.063)
Additional covariates	Yes	Yes	Yes	Yes
Observations	225	225	135	135
Clusters	5	5	5	5

Note: The table reports probit regression estimates of the determinants of the PEER EVAL vote for best team member. Columns (1) and (3) report marginal effects from probit regressions where each variable of interest is included separately while columns (2) and (4) include all variables in the same regression. The sample consists of all subjects who received the PEER EVAL treatment minus 9 observations due to technical problems with the video. Additional covariates consist of the standard set of individual controls. Robust standard errors clustered by session are reported in parentheses. P-values stem from a wild bootstrap correction (STATA boottest). Asterisk indicate significance at the 10/5/1 percent level.

individuals with "Different" answer sheets to more often present their whole method, including the creative steps, to the group. In general, as Table 9 shows, an individual with a "Different" individual answer sheet had a higher chance of presenting the whole method and also presenting the whole method as the first in the group. Furthermore, this difference is more pronounced under PEER EVAL than under the other treatments consistent with the peer evaluation rewarding the more easily observable details of the solution approach, such as the originality of the steps, instead of the quality of the final answer.

Table 9: Propensity to Present Method by Different

		FLAT	GROUP INC	PEER EVAL
Method presented	Different indiv. sheet	0.43	0.46	0.52
	Normal indiv. sheet	0.36	0.39	0.34
	Difference	0.07	0.07	0.18
Method presented first	Different indiv. sheet	0.29	0.31	0.35
	Normal indiv. sheet	0.19	0.18	0.16
	Difference	0.09	0.12	0.18

Some indication that this was indeed the case can be found in the answers in the questionnaire, where we asked subjects about the reason for their vote. While this is clearly nothing more than anecdotal evidence, some of these answers hint at the originality of steps mattering:

"He thought about the situation in a different way and had a reasonable answer."

"Both were really good, but he had a few good ideas such as the 06-06-06 bonus."

"More simple logic, good innovative ideas, structured thinking."

Together with the fact that the best of the three individual guesses was on average better than the guess that the group ultimately decided on (16.1 euro vs. 12.8 euro in terms of hypothetical group payoff, whole sample) the results seem to suggest that the peer evaluation was an imperfect incentive mechanism also because subjects were not able to easily recognize (and improve) the best individual guess.[26] While groups were not able to fully make use of the best individual guess,

[26]While the best individual guess was superior, the group guess was significantly better than the second-best guess

individual guess quality still mattered. The correlation between the best individual guess payoff and the group guess payoff equals 0.48 indicating that better individual guesses did translate into better group guesses on average. As it seemed difficult to assess the quality of a guess directly, subjects may have based their judgement on more easily observable, imperfect and indirect, measures of performance. If the link between guess quality and these measures was not strong enough, this may explain why the peer evaluation, while effective in increasing motivation and effort, ultimately did not lead to better performance.

To sum up, we found no significant effect of incentives on aggregate performance, though we did observe that motivation and effort increased when a peer evaluation was present. Changes in behavior are consistent with the peer evaluation rewarding the details and originality of the solution method and an individual's behavior during the group phase without regard for the quality of the individual answer. Group incentives without peer evaluation had a small effect on group effort and no effect on motivation and performance.

4 Gender Differences

In this section we present the results of an exploratory analysis of whether males and females respond differently to incentives. Since we did not design the experiment with a hypothesis on gender differences in mind, we first present our findings, and then relate them to the findings in the literature. In the following, we summarize the analysis conducted in Section 3.2 distinguishing between males and females at the individual level and male-majority (0 or 1 female) and female-majority (2 or 3 females) groups at the group level.[27]

Examining the ability guesstimation, females earned on average 2.95 euro, while males earned

(6.3 euro) or a random guess (8.9 euro) (whole sample). Groups also added value relative to the average (arithmetic and geometric) or median of group members' guesses. Thus the group phase is valuable to a principal. Table 12 in the appendix summarizes the relationship between individual and group guess quality for groups where at least two individuals submitted an individual guess.

[27]While subjects were allocated to groups randomly, and thus differences in group composition are exogenous to treatment, there were overall less females in the subject pool (38%). Thus we have few female-only groups. For this reason we constrain ourselves to a more aggregated analysis of gender effects and pool groups with no or one female into a "majority male" category, and groups with 2 or 3 females into a "majority female" category. We control for the number of females in all group-level regressions. Table 13 in the Appendix shows group performance (hypothetical payoffs) by number of females for each treatment as well as the number of observations for each case.

3.47 euro out of a maximum of 10 euro. This difference is significant at the 5% level (ranksum test, p-value 0.031). Since females performed worse at the individual ability task, group differences in performance may be driven by gender composition directly. But, conditional on ability, males and females may also be reacting to incentives differently. We therefore compare treatment effects by group composition. Table 10 shows the regression results of the main treatment effects in Table 4 of the previous section by majority gender, focusing on the hypothetical payoff performance measure. We find that female-majority groups performed significantly worse in PEER EVAL, relative to GROUP INC (the difference is not significant though of the same sign relative to FLAT), while for male-majority groups the effect is positive but insignificant. Also economically the effect is quite large: female-majority teams earned between 5 and 7 euro less in PEER EVAL relative to GROUP INC depending on the specification (between 15% and 20% of the maximal earnings of 35 euro).[28][29]

Differential effects of the PEER EVAL treatment may stem from males and females being differently evaluated in the peer evaluation. For example, if evaluations are biased against females, this may demotivate them. We do not find evidence of this in our setting. Males and females had the exact same propensity to be voted winner: 28%[30]. Also regression analysis reveals no significant difference between males and females in their propensity of winning the peer evaluation controlling for observable behavioral differences, such as their speaking turns share as well as individual guess quality. The wild bootstrap corrected p-values of the coefficients on female in Table 8 column (2) and (4) equal $p = 0.3125$ and $p = 0.7500$ respectively.

Next we explore whether females and males were affected differentially in their motivation to exert effort. Consider the first 5 minutes of the group task where subjects filled out their individual answer sheets. One of the largest treatment differences found in the previous section in Table 5 were the high numbers of missing individual guesses in the PEER EVAL treatment. Table 14, in the Appendix, which disaggregates Table 5 by gender, shows that this difference is driven by

[28]The results for the more continuous (but also more noisy) performance measure, percentage error, are insignificant though qualitatively similar.

[29]As a robustness check, we also perform variable selection through a linear Lasso analysis. The variables selected by the Lasso are: an indicator for one female group member, the interaction between the treatment indicator for the peer evaluation and the indicator for majority female groups (GI&PE*Majority female) as well as the number of econometrics students in the group. This confirms the importance of gender for the treatment effect of the peer evaluation.

[30]Note that in some groups each group member received one vote and thus there is no winner in that group.

Table 10: Heterogeneity in Treatments Effects on Group Performance (Hypothetical Payoffs) by Group Gender Composition

	Majority male		Majority female		All	
	(1)	(2)	(3)	(4)	(5)	(6)
Group Incentives	0.186	0.296	2.504	5.308	0.186	-0.658
	(3.324)	(3.396)	(2.791)	(4.066)	(3.339)	(3.230)
Group Inc. & Peer Evaluation	2.907	2.250	-5.060**	-6.976***	2.907	2.702
	(1.831)	(2.601)	(2.210)	(2.788)	(1.840)	(2.476)
Group Inc.*Majority female					2.318	3.535
					(4.457)	(4.936)
GI&PE*Majority female					-7.967**	-7.623**
					(2.958)	(2.915)
Majority Female					-0.111	-1.051
					(3.884)	(4.531)
Additional Covariates	No	Yes	No	Yes	No	Yes
Observations	160	160	71	71	231	231
Clusters	15	15	14	14	15	15

Note: The table reports OLS estimates of the impact of group gender composition on group payoff. The first two columns consider groups with 0 or 1 females, the second two with 2 or 3 females. Robust standard errors clustered on the session level are reported in parentheses. Additional covariates consist of the mean and maximum group ability score, a measure for group SVO, measures for group Big5 items, number of females, the average age of group members, group same nationality, percentage first and second year bachelor students, number of people with experience with the task, the number of Economics students, the number of Econometrics students and average as well as maximum grade point average. Asterisks indicate significance at the 10/5/1 percent level. P-values are calculated using wild bootstrap (STATA boottest) to account for the small number of clusters.

females. For females the number of missing guesses doubled between GROUP INC and PEER EVAL, a significant increase, while for males it stayed constant. In addition, the guesses that were submitted by females in the individual phase yielded only half the payoff under PEER EVAL than under GROUP INC (10.41 euro vs. 5.7 euro, though this difference does not quite reach statistical significance) while we do not find any such difference for males. On the other hand, there was a significant increase in the number of steps in male answer sheets which did not contain a final guess between GROUP INC and PEER EVAL.

Turning to the group phase, Table 15, in the Appendix, shows Table 6 split up by majority group gender. There were no systematic gender differences in the number of guesses/methods presented though it does seem that in female-majority groups less information was exchanged in general, as found also for example in Coffman (2014). Overall results for male- and female-majority groups were similar for measures of group effort: work time, conversation about unrelated matters as well as speaking turns. Effort increased with incentive strength, though the increase in effort seemed to happen more pronounced going from FLAT to GROUP INC for female-majority teams, while for male-majority teams it was more pronounced going from GROUP INC to PEER EVAL. We find no significant differences for either type of group in terms of reaching a final agreement or the chance of dominant individuals.

Finally, in the questionnaire on group atmosphere and motivation, we study the answer to the statement "I wanted to make a good impression on my group members" as a proxy for motivation by gender. While Table 7 shows that under PEER EVAL subjects agreed more with this statement, this may be driven by male subjects. This turns out not to be the case. Both males and females were significantly more in agreement with the statement under PEER EVAL relative to the other treatments. Tables 16 and 17 in the Appendix show Table 7 split up by gender of the respondent. In general, we observe similar patterns for both genders in their answers to the various statements. Thus it seems that females as well as males were more motivated and increased their effort in the PEER EVAL treatment.

Summarizing, while both males and females were motivated to exert more effort in PEER EVAL, and seemingly also did so in the group phase, females entered the group phase with less (and worse)

individual guesses and female-majority groups ended up performing worse, in particular compared to GROUP INC. One explanation identified in the literature that is consistent with these findings is that females are choking under the pressure of the peer evaluation. Ariely et al. (2009) show that high-stakes incentive pay may cause choking under pressure. In their setting, they do not find gender effects. Other studies do find differential effects for females and males, such as Bracha and Fershtman (2013) and Cahlíková et al. (2019), in a laboratory setting, as well as Azmat et al. (2016) and Cai et al. (2019) for high school test takers.[31]

In a final attempt to explore whether choking under pressure by females rationalizes our data, we relate the answer to the question "I wanted to make a good impression on my group members" to individual performance after the first 5 minutes for those females and males that did submit an individual guess in the PEER EVAL treatment as well as the propensity to not submit an individual guess. Table 11 summarizes the results. The results are consistent with choking under pressure for females — the degree of agreement to the "good impression" question is negatively related to individual performance (conditional on submitting an individual guess). For males we observe the opposite. Performance is positively related to motivation as measured by the question. For missing guesses we see no clear relationship for females, while males who state a higher motivation were more likely to have an individual guess.

Table 11: Individual performance of subjects that wrote down a final individual answer as well as fraction of missing guesses by gender and motivation in the PEER EVAL treatment. The number of observations is stated in brackets.

"I wanted to make a good impression on my group members"					
	Disagree		Neither	Agree	
	Fully	Mostly		Mostly	Fully
Male individual payoff	n.a.	4.7 (3)	7.6 (12)	10.5 (48)	11.4 (46)
Female individual payoff	n.a.	n.a.	8.2 (12)	6 (42)	3.1 (16)
Male missing guess	n.a.	0 (3)	.4 (20)	.19 (59)	.08 (50)
Female missing guess	n.a.	1 (1)	.33 (18)	.28 (58)	.36 (25)

[31] Some studies also find males to be more affected by choking under pressure, such as Cohen-Zada et al. (2017) for professional tennis players.

5 Conclusion

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 the face of differing incentive schemes. We compare performance under fixed pay, pay that depends on group performance as well as pay that in addition depends on the outcome of a peer evaluation.

Overall, we find no significant effect of any type of incentive on performance. While having no incentives and having group incentives did not produce very different group outcomes, groups behaved significantly differently under a peer evaluation. Subjects reported higher motivation and groups worked harder and longer in the group phase. This seemed to stem from the peer evaluation being an imperfect incentive mechanism: success in the peer evaluation was not related to individual performance but instead it was positively related with easily observable proxies of effort, such as speaking turns and number of steps. An interesting question for future research is whether the bias identified persists even if subjects collect more experience with the task.

Finally, our results of an exploratory analysis point towards a differential gender effect in the response to a peer evaluation. While male-dominated teams' performance was largely unaffected, female-dominated teams performed relatively worse than in the absence of a peer evaluation. We leave it for further research to more carefully study the causes and consequences of such a differential effect and explore whether modern approaches to performance evaluation systems based on peer evaluation may be contributing to a gender gap on the labor market.

What does this mean? The idea of crowdsourcing rating and review is, by now, very familiar. From Digg to Reddit to the "like" button on Facebook, the success or failure of internet items is collectively determined. Inside the corporation, 360 assessments, evaluations of an individual by boss, peers, and underlings, are becoming more common and influential in career rewards. Our results, however, suggest that individuals are quite poor at recognizing, and rewarding, good solutions to problems. At least for mainly intellective tasks, we advise caution for the use of peer-based systems, such as 360 assessments.

6 Appendix

6.1 Supplementary Tables

Group vs. Individual Guesses

Table 12: Mean Group and Individual Payoffs by Question

	Cycling		**Toothpaste**		**Weddings**		**Pooled**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group								
Payoff	13.90	11.131	13.06	11.553	12.97	12.814	13.32	11.795
Individual								
Best Payoff	17.85***	11.580	16.51**	11.921	17.60***	12.655	17.33***	12.009
Second Best Payoff	9.17***	9.509	4.39***	6.983	5.35***	8.389	6.35***	8.598
Random Payoff	10.66**	8.057	8.04***	6.569	9.08***	7.571	9.29***	7.481
Measure of Central Tendency								
Arithmetic Mean	11.14*	11.925	8.25***	10.962	7.51***	11.620	9.00***	11.571
Geometric Mean	14.59	11.638	10.55	12.208	8.96**	12.501	11.42**	12.290
Locational Median	14.89	11.769	10.45*	11.193	9.99*	11.992	11.83*	11.816
# Observations	71		67		68		206	

Note: Only groups where at least two individuals made a guess are included. Having made a guess is ranked better than no guess. Random Payoff is calculated as the mean of the individual payoffs. Arithmetic (Geometric) Mean Payoff is calculated as the payoff corresponding to the arithmetic (geometric) mean of the individual guesses. Locational Median Payoff is calculated as the payoff corresponding to the second highest guess if there were three guesses or the geometric mean of the guesses if there were two. Significance stars are added to the individual payoffs, indicating the p-value of a paired t-test for the two sided hypothesis: *group payoff* $\neq$ *individual/mean payoff*. Significance levels are denoted as follows: * p$<$ 0.1, ** p$<$ 0.05, *** p$<$ 0.01.

Gender Differences

Table 13: Group Payoffs by Group Gender Composition and Treatment.

	Majority male		Majority female	
	0 females	1 female	2 females	3 females
FLAT	**12.17 (42)**		**12.06 (18)**	
	13.53 (15)	11.41 (27)	12.6 (15)	9.33 (3)
GROUP INC	**12.35 (68)**		**14.56 (25)**	
	11.15 (27)	13.15 (41)	13.39 (23)	28 (2)
PEER EVAL	**15.26 (32)**		**9.5 (28)**	
	8.08 (13)	17.78 (27)	11.79 (19)	4.67 (9)

Table 14: Treatment Differences Individual Phase by Gender

		FLAT	GROUP INC	PEER EVAL	
Missing Guesses	Males	0.09	0.18	0.17	
	Females	0.06	0.14	0.31***	**
Individual Payoff (complete)	Males	9.02	10.18	10.40	
	Females	7.34	10.41	5.70	
Ind. Steps (all)	Males	5.17	4.88	4.89	
	Females	4.61	4.72	4.70	
Ind. Steps (complete)	Males	5.26	5.06	4.84	
	Females	4.61	4.81	4.86	
Ind. Steps (missing)	Males	4.20	4.09	5.09*	*
	Females	4.50	4.15	4.34	
Different	Males	0.08	0.11	0.11	
	Females	0.11	0.06	0.10	

Note: The table reports group means. Asterisks directly after a statistic indicate a difference of means compared to FLAT significant at the 10/5/1 percent level. The last column compares GROUP INC and PEER EVAL. The p-values for payoff and steps are calculated from a regression clustering standard errors at the session level corrected using wild bootstrap (STATA boottest). The p-values for different and missing guess are calculated from a probit clustering standard errors at the session level corrected using wild bootstrap (STATA boottest). The standard set of group control variables is included in all specifications.

Table 15: Treatment Differences Group Phase by Majority Group Gender

		FLAT	GROUP INC	PEER EVAL	
Guesses Shared	Majority Female	1.81	1.08*	1.23*	
	Majority Male	1.88	1.44*	1.63	
Methods Shared	Majority Female	0.88	0.88	1.00	
	Majority Male	1.22	1.32	1.12	
Worktime	Majority Female	7.34	8.66*	8.88	**
	Majority Male	8.29	8.25	8.81***	**
Unrelated	Majority Female	0.69	0.25***	0.46**	***
	Majority Male	0.69	0.59*	0.51*	
Number of Speaking Turns	Majority Female	8.69	10.04	11.19	
	Majority Male	10.22	10.60	12.27	
All Agree	Majority Female	0.88	0.79	0.77	
	Majority Male	0.88	0.89	0.86	
No Dominant Indiv.	Majority Female	0.44	0.13	0.42	
	Majority Male	0.34	0.41	0.35	
Different	Majority Female	0.00	0.00	0.11	
	Majority Male	0.07	0.06	0.08	

Note: The table reports group means. Asterisks indicate a difference of means compared to FLAT significant at the 10/5/1 percent level. The last column compares GROUP INC and PEER EVAL. The p-values for guess, method, worktime and turns are calculated from a regression using robust standard errors clustered at the session level corrected using wild bootstrap (STATA boottest). The p-values for unrelated, all agree, dominant are calculated from a probit regression using robust standard errors clustered at the session level corrected using wild bootstrap (STATA boottest). Due to the low number of different answer sheets we are not able to report p-values for different. The standard set of group control variables is included in all specifications.

Table 16: Perception of Group Environment by Treatment - Females

	FLAT	GROUP INC	PEER EVAL	
I wanted to make a good impression on my group members	3.515	3.591	4.049**	***
The atmosphere in the group was helpful	4.242	4.065**	4.049*	
The atmosphere in the group was competitive	2.227	2.172	2.706**	***
Do you feel that competitiveness helped the group reach a better performance	3.382	3.137	2.941*	
I felt that everyone had an opportunity to voice their ideas in a fair way	4.576	4.419*	4.324**	
I felt that others dominated the discussion in unproductive ways	1.803	1.656	1.961	**

Note: Answers to the survey questions range from 1 fully agree - 5 fully disagree. The p-value are obtained from an ordered logit regression clustering standard errors at the session level corrected using wild bootstrap (STATA boottest). The standard set of individual control variables is included in all specifications. Significance stars are shown as follows: * p< 0.1, ** p< 0.05, *** p< 0.01. Asterisks indicate a difference of means compared to FLAT significant at the 10/5/1 percent level. The last column compares GROUP INC and PEER EVAL. Note that the number of observations are lower for the 4th question because a percentage of the subjects who didn't perceive the group as competitive skipped this question.

Table 17: Perception of Group Environment by Treatment - Males

	FLAT	GROUP INC	PEER EVAL	
I wanted to make a good impression on my group members	3.500	3.661	4.182***	***
The atmosphere in the group was helpful	4.114	4.102	4.242	
The atmosphere in the group was competitive	2.447	2.349	2.720	***
Do you feel that competitiveness helped the group reach a better performance	3.013	3.339	3.287	
I felt that everyone had an opportunity to voice their ideas in a fair way	4.412	4.435	4.417	
I felt that others dominated the discussion in unproductive ways	1.974	1.860	2.114	*

Note: Answers to the survey questions range from 1 fully agree - 5 fully disagree. The p-value are obtained from an ordered logit regression clustering standard errors at the session level corrected using wild bootstrap (STATA boottest). The standard set of individual control variables is included in all specifications. Significance stars are shown as follows: * p< 0.1, ** p< 0.05, *** p< 0.01. Asterisks indicate a difference of means compared to FLAT significant at the 10/5/1 percent level. The last column compares GROUP INC and PEER EVAL. Note that the number of observations are lower for the 4th question because a percentage of the subjects who didn't perceive the group as competitive skipped this question.

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*, 51:979990.

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. (2017). Do women respond less to performance pay? Building evidence from multiple experiments. working paper.

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

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.

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*, forthcoming.

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.

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. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):16251660.

Coffman, K. B., Flikkema, C. B., and Shurchkov, O. (2019). Gender Stereotypes in Deliberation and Team Decisions. Harvard Business School Working Paper, No. 19-069.

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.

Deb, J., Li, J., and Mukherjee, A. (2016). Relational contracts with subjective peer evaluations. *The RAND Journal of Economics*, 47(1):3–28.

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):627668.

Delfgaauw, J., Dur, R., and Souverijn, M. (2017). Team incentives, task assignment, and performance: A field experiment. Tinbergen Institute Discussion Paper 17-090/VII.

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.

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.

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

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.

Kim, J.-H. (2011). Peer performance evaluation: Information aggregation approach. *Journal of Economics & Management Strategy*, 20(2):565–587.

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

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

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.

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

Sol, J. (2016). Peer evaluation: Incentives and coworker relations. *Journal of Economics & Management Strategy*, 25(1):56–76.

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.

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.